Parametrising kinetic models of biological networks

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Dr. rer. nat.
im Fach Biophysik/Theoretische Biophysik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Dipl.-Phys. Simon Borger
geb. am 30.03.1975 in Berlin

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Christoph Markschies

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Dr. Christian Limberg

Gutachter:
1. Prof. Dr. Dr. h.c. Edda Klipp
2. Prof. Dr. Hermann-Georg Holzhütter
3. Prof. Dr. Thomas Höfer
4. Prof. Dr. Matthias Reuss

eingereicht am: 28. Januar 2008
Tag der mündlichen Prüfung: 10. Dezember 2008

Abstract

Systems biology seeks to model biological networks dynamically. Two requirements need to be fulfilled for this to be possible. First, the interaction networks need to be known. Second, the dynamics of the interactions have to be revealed. Dynamics of interactions are described by rate laws using kinetics. These kinetics describe the interaction mechanism. For each single interaction occurring in a biological network parameters have to specified. They have to be measured by experiments. For enzyme catalysed reactions, for example, the parameters are measured by enzyme assays tracking the consumption of substrate. For many enzymes parameters and kinetic mechanism are not known. And existing data for parameters generally are of poor quality. After introducing kinetic modelling of metabolic networks we consider a published artificial genetic network that can either tend to a steady state or exhibit sustained oscillations depending on a critical parameter. This critical parameter is the Hill coefficient in the interaction from one gene with the other. For different parameter settings we examine at what value of the Hill coefficient a bifurcation occurs. At this point the network begins to oscillate. We thus assess the distribution of the critical values, a property that cannot be calculated analytically. We then go on to consider useful data sources for parmetrisation of kinetic models of metabolic networks and collect them in an electronical resource to make them electronically accessible and usable. This requires choosing standard references for the designation of components of biological networks. Finally we describe a workflow in which this data resource is used for automatic parametrisation of kinetic models of metabolic networks.

Zusammenfassung

Systembiologie strebt danach, biologische Netzwerke dynamisch zu modellieren. Zwei Erfordernisse sind zu erfüllen, damit dies möglich wird. Erstens müssen die Interaktionsnetzwerke bekannt sein. Zweitens muss die Dynamik einer jeden Interaktion aufgedeckt werden. Die Dynamik von Interaktionen werden durch Ratengleichungen beschrieben unter Verwendung von Kinetiken. Diese Kinetiken beschreiben den Interaktionsmechanismus. Für jede einzelne Interaktion, die in einem Netwerk auftritt, müssen Parameter bestimmt werden. Sie sind durch das Experiment zu messen. Für enzymkatalysierte Reaktionen zum Beispiel werden Messungen durchgeführt, in welchen der Verbrauch des Substrates aufgezeichnet wird. Für viele Enzyme jedoch sind weder der Mechanismus geschweige denn die Parameter bekannt. Und vorhandene Daten sind gewöhnlich von mangelhafter Qualität. Nach einer Einführung in die kinetische Modellierung metabolischer Netzwerke betrachten wir ein veröffentlichtes künstliches genetisches Netzwerk, das entweder einem stationären Zustand zustrebt oder in Abhängigkeit eines kritschen Parameters in einen dauerhaften Schwingungszustand übergeht. Dieser kritsche Parameter ist der Hillkoeffizient in der Wechselwirkung zwischen einem Gen und dem anderen. Für verschiedene Parameterwahlen untersuchen wir, bei welchem Wert des Hillkoeffizienten eine Bifurkation auftritt. Hier fängt das Netzwerk zu schwingen an. Auf diese Weise ermitteln wir die Verteilung des kritschen Parameters, der nicht analytisch berechnet werden kann. Wir fahren dann fort und untersuchen nützliche Datenquellen für die Parametrisierung von kinetischen Modellen metabolischer Netzwerke und sammeln sie in einer elektronischen Ressource, um sie auf elektronischem Wege zugänglich und nutzbar zu machen. Dies erfordert, Standardreferenzen zu wählen für die Benennung der Komponenten biologischer Netzwerke. Schließlich beschreiben wir einen Arbeitsablauf, während desselben die Datenbank verwendet wird zur Parametrisierung von kinetischen Modellen metabolischer Netzwerke.

Contents

1 Systems biology

1.1 Life is more than the sum of its parts

The ultimate goal of biology is to understand life. The domain of life is the world of organisms as opposed to the inorganic, dead matter. Living organisms are characterised by growth, reproduction and adaptation to their environment. All species from the unicellular to multicellular creatures, like mammals with intricate anatomies, live and survive in an environment that at the same time is friend and foe, that feeds and menaces them while they go through their life cycle producing offspring. In the process of reproduction the genetic material is passed on to the offspring to equip them with the arsenal that enables them to survive and proliferate.

Underlying this passage through the life cycle is a wealth of processes. These processes comprise uptake and digestion of nutrients, defense of stressors, the assembly of cellular components, processing of and response to hormones. They are at the basis of growth and reproduction and give rise to what is called the *functions* of an organism, i.e. all those means and ways the organism shows up with at the physiological level to make its species survive.

All these processes are achieved by thousands of proteins cooperating to form functional modules. Among these are signalling pathways, metabolic pathways and the gene expression apparatus. Proteins are important in the cell's signal transduction apparatus for immune responses or cell fate decisions. They drive the cell cycle or facilitate metabolism by speeding up biochemical reactions. The cell's shape is supported by proteins forming the cytoskeleton. Proteins are also involved in muscle contraction, cell division and motility.

Molecular biology and biochemistry have revealed a lot about the basic processes of life. Genes were discovered and the helical structure of the DNA was determined. How genes are induced by activators or impeded by repressors has been described. Proteins have been sequenced and exposed to X-rays to reveal their complex three dimensional shape. That differences in morphogen are decisive in cell development has been understood.

But life hides more puzzles than those uncovered by molecular biology. An example is robustness which is the property of a system to maintain its function and pursue its purpose despite external or internal perturbations. For instance the bacterium *Escherichia coli* living in the intestines of humans shows this remarkable property. It directs its movement towards a source of nutrient by sensing where the nutrient concentration is highest in its surrounding. Amazingly, this sensory system performs reliably across a wide range from small to large concentration differences. Robustness is a phenomenon encountered in biology that cannot be explained on a molecular basis going through a

series of events step-by-step.

Recent years have seen the advent of a new era in biology. This era is witnessing the breakthrough of a new field called systems biology. Although lacking a solid definition this name has generally been accepted and is widely used. Sometimes systems biology is characterised as being the opposite to the reductionist approach, rather taking a holistic standpoint and considering entire biological systems or subsystems than dissecting them into molecules. Systems biology is an integrative discipline that builds on the revealed molecular details and takes a step back to regard the system as a whole.

In 1961, W.B. Astbury writes that he perceived molecular biology, generally viewed as an embodiment of reductionism, as "an approach from the viewpoint of the so-called basic sciences with the leading idea of searching below the large-scale manifestations of classical biology for the corresponding molecular plan." [9] Molecular biology has studied the components of the molecular plan of the cell in isolation. The belief of molecular biologists was that an appropriate description of the large-scale properties of the cell would arise from simply joining the properties of its components into a large picture. That this view is a too reductionist is being affirmed by a growing number of biologists.

So systems biology then is concerned with those large-scale manifestations. It might seem that biologists are walking around in circles, leaving those large-scale manifestations of classical biology for the reductionist and more basic perspective of molecular biology to finally return to the large-scale manifestations with systems biology.

In its so-called bottom-up approach, systems biology unites knowledge about components of the large-scale manifestations and their interactions. "It is about putting together rather than taking apart, integration rather than reduction. It starts with what we have learned from the reductionist approach; and then goes further. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different. This is a major change. It has implications beyond the purely scientific. It means changing our philosophy, in the full sense of the term." [111].

Once the components have been put together to form a system, how then does it behave? Systems biology is setting out to explain the dynamical properties of these systems based on the properties of their parts and how they act upon each other. The dynamics are brought about by the interactions among the parts. In the focus systems biology is not only the question of which components interact, but also to what extent they interact and how the interactions change during the course of time.

The dynamic properties that *emerge* from the systems level description of the components and their interactions are the novel point about systems biology. It is not possible to understand the properties of systems and thereby their *functions*, without examining *how* their parts are connected and drive each other. One does not know what a leg is good for if one disregards the the rest of the body it is attached to. An ant from the worker caste fulfills its function only as a cogwheel within the whole colony.

Examples of *emergent* system properties in the cell are memory or periodic rhythms. Memory emerges as a property when a system being in either of two states cannot effortlessly be prompted to move to the other. This phenomenon is called bistability. The memory of the brain or the immune system and cell fate decision are explained based on bistability [147, 17, 89]. In a model of the cell cycle the critical cell volume on which

onset of DNA replication depends, emerges as a system property [11]. A model of the cardiac muscle brings about the pacemaker rhythm of the heart [110]. Oscillations are also seen in engineered genetic regulatory networks as consequence of negative feedback like in a cycle with successive inhibitions [45].

By revealing basic functions of living organisms by studying systems as whole, biology is walking in circles and coming back to the large-scale manifestations in the sense that the perspective is a holistic one. But this approach is based on the reductionist view. The systems studied in systems biology are made up of the parts that molecular biology unearthed. And systems biology now is putting the puzzle together and discovering that living systems are more than the sum of their parts.

1.2 Living networks

The objects of study of systems biology are biological *networks*. A network is per definition a collection of nodes and of connections between them. The nodes communicate through the connections. The message communicated depends on the state of the interacting nodes, on the type of interaction and possibly on external factors.

A key feature of networks is their dynamics. That is, the state of a network undergoes temporal change because the nodes interact. Often, without external forces a network will tend towards a constant state.

Everyday life examples of networks are landline or cellular telephone networks, computer networks like the internet, social networks like bridge clubs, or traffic networks like a railway system.

If we take a closer look at the rail network example, we could ask ourselves what the railway operator has to consider in order to secure good service of the rail network. For instance there should be no time delays and thus no hold up on the tracks.

In a rail network trains travel constantly between different destinations transporting people and goods.

Good functioning requires not only a timetable such that trains are coordinated, but also measures against unforeseen incidents like a train that has jumped out of the rails or damages through weather. A rail network requires overall planning and enough flexibility to be able to maintain a minimum of service in unforeseen cases.

In biology networks are ubiquitous. The brain is a network of neurons that exchange signals through synaptic transmission. The food chains of animals and plants constitute ecological or predator-prey networks. Any higher-level creature contains a hierarchy of networks. The organs and the blood system constitute a network. Tissues consist of many interacting cells. And finally, the cell itself is a complex network.

The cellular network is far from being fully apprehended. The cell is a network that can only be understood in terms of its dynamics arising from the interactions. The central dogma of molecular biology, stating that a gene is transcribed and translated into a protein, brings the two main networks into play. A cell's genes and its proteins are the key players of a living cell. The genes contain the building plans of the proteins. The building plans entail certain attributes of the proteins that make them acomplish

certain purposes within the cell. They assist in turning sugar into energy or in fighting and disarming malign intruders. Coordinating all those proteins lets the cell fulfil its functions like growth, immune response or cell devision in a coordinated fashion to finally reproduce.

Biological ones are built of nodes and interactions between them. The set of all genes, the genome, form a network. A gene is expressed, through transcription and translation, to bring forth the coded protein. Either directly, as a transcription factor, or indirectly, in a signalling pathway, the protein then acts back on the genome, possibly the original gene or others. The genes form a network and interact through the proteins they encode.

From a different perspective, the set of proteins make up an interaction network. One can distinguish between different types of protein networks depending on the nature of the interaction. They could be called the cellular computer, the cellular factory and cellular workhorses, respectively.

The networks that make up the cell's computer are the so-called signalling pathways. The cell receives signals from its environment and processes them by signalling cascades. Proteins activate and inactivate each other, switch each other on and off, by exchanging functional groups like phosphate. At the end of the cascade a protein is switched on and affects the transcription of target genes. By thus processing signals the cell reacts to changes in the environment. Hormones for instance invoke cellular differentiation during embryo development to grow different tissues and parts of the anatomy.

The cellular factory consists of another group of proteins, the enzymes. Nutrients like sugar are digested for energy production that is needed in other cellular processes like the synthesis of the proteins. The sugar gets broken down into smaller molecules through a series of biochemical reactions, the so-called metabolic pathways. Enzymes drive metabolism by speeding up the biochemcial reactions. Enzymes interact indirectly by passing down metabolites from one to the other along the metabolic pathways.

The third type of protein interaction network, the cellular workhorses, are single large molecules composed of many proteins bound in a complex. The proteins fulfil their purpose only in conjunction with the others in the complex. The binding of the proteins is stable over time as compared to those protein-protein interactions in signalling pathways where proteins bind shortly to exchange groups. An example of such a molecule requiring many proteins bound in a complex is the ribosome. It consists of RNA and several tens of proteins. The ribosome is responsible for protein biosynthesis in gene expression by translating the genetic code into an amino acid sequence.

The interactions among the components of the cell that make up the cellular networks and govern its dynamics are intrinsically non-linear. This means that the rates of processes in the cell do not grow or shrink proportionally to what drives them, i.e. the abundances of the biomolecules. A well known non-linearity is found in the rate of enzyme-catalysed reactions, the rates of which are limited by the total enzyme amount, leading to saturation and a maximal reaction rate. Other non-linearities in many metabolic pathways and gene regulatory networks arise from feedback. In negative feedback, for instance, the response decreases a signal after an increase of the signal. An example is the inhibition of the glucose transport into the cell by Glucose 6-phosphate, a metabolite in glycolysis pathway, after uptake of glucose.

And indeed, the non-linearities of the interactions are at the core of control and regulation of cellular processes. No cell could survive an ever increasing concentration of a metabolite. So there have to be means and ways to keep the cellular state within a physiological range despite of environmental influences. Some properties of organism even stay constant in changing environments like the body temperature of mammals.

The non-linear interactions give rise to the cell's functions. This is what is meant when one speaks of properties of networks emerging from the interactions. These emerging properties are often said to be counter-intuitive [17, 138] which is due to the non-linear nature.

For instance, protein networks involved in signalling or the control of metabolism exhibit ultrasensitivity and constitute switches swapping between two states. Due to the non-linearities the switch can be very sharp with an on-off-profile requiring a signal above a threshold to get the response [65, 66]. Memory arises in cell development from positive feed-back, a non-linear interaction causing two stable states for a certain range of signal strength and discontinuous switching at the borders of this range [138]. The value of the response within the signal range of bistability depends on the history of the system. The adaptation of organisms to the daily rhythm, so-called circadian rhythms, also emerge from non-linear feedback properties of the network interactions [64].

1.3 Models of biological networks

Setting up models of biological interaction networks like a metabolic or a signalling pathway is a three-fold process. As networks are described by nodes and interactions, first the nodes and the connections between them have to be identified.

Second, the nature of the interaction between two nodes has to be assessed. This is especially true for enzymes when several metabolites and effectors can bind to it. For instance do we need to know for an enzyme catalysing a reaction with two substrates whether the temporal order of binding to the enzyme is relevant. There might also be effectors inhibiting or activating the binding of a metabolite.

Third, the kinetic expressions of the interactions need to be assigned parameters. The mathematical description of the non-linear interactions in the cellular networks phenomenologically sums up detailed processes. Protein-protein binding or substrate-enzyme binding are intricate series of events involving conformational changes of the proteins. Mass action or kinetics derived therefrom are coarse-grained views of these events encompassing different time scales. Different kinetics for protein binding have been developed from mass action by separating fast from slower processes, among them the famous Michaelis-Menten and the Hill kinetics. As these kinetics are phenomenological descriptions of detailed processes the parameters for the kinetics describing them need to be measured for every single possible interaction.

For what concerns the first step, the identification of the cellular protein networks, molecular biology, biochemistry and genetics have set out a picture of the hierarchy of networks and its components. As the catalogue of components is so tremendous this job is still not finished. In the nineties the rise of genomics led to the decoding of

the first genomes containing millions of base pairs and culminated in the completion of deciphering the euchromatic human genome in 2004 [34]. In the following years then, between 1995 and 2004, the number of deciphered genomes grew exponentially [121]. As of October 2007 the number of complete genomes, publicly available at NCBI (National Center of Biotechnolgy Information), is 24 for eukaryotes [51] and 594 for archaea and bacteria [52]. The sequencing of complete genomes has accelerated the discovery of new genes by 50–100 times [118] compared to classical genetics.

To date the function of many genes and their protein products remains unknown. In 1996, the year the genome of the first eukaryote, of the yeast *Saccharomyces cerevisiae*, was completely deciphered, nearly half of the genes newly discovered through genome sequencing was completely unknown [118].

The knowledge of the mere genomic sequences is of value in evolutionary and phylogenetic studies as their similarity within and across species can be studied. Apart from that, the genetic sequence only determines the amino acid sequence of the encoded protein and tells us nothing about the process and function the protein might play a role in (despite efforts to explain protein properties from their sequences). So for systems biology, sequenced genomes alone remain meaningless series of the letters A, C, T and G like the words of a foreign language the words of which we cannot convert into sound or meaning.

To get a feeling of the effort that awaits biologists when it comes to make meaning of the human proteome take a look at the size of the genome. It has a length of approximately 3000 Mbp. It is 250 times larger in terms of nucleotides than the genome of *Saccharomyces cerevisiae*. There had been wild speculations about the number of human genes contained in those 3000 Mbp. By the year 2004 it came as a big surprise that the human genome contains only 20000–25000 genes [34], that is only approximately 4 times as many as for *Saccharomyces cerevisiae*. But this number of genes is faced by a much larger number of proteins due to a combinatorically increased number of ways of reading the single gene through alternative splicing and post-translational modification [82, 20]. Current estimates say that one gene can yield up to 60 different proteins leading to a number of different human protein species that exceeds one million [82]. At the UniProt database [6] in the current version 17316 human proteins have reviewed annotations [37] which amounts to on the order 1% of the protein total.

What about the functional annotation of the genome of a smaller and well studied organism? The yeast *Saccharomyces cerevisiae* is a so-called model organism being the first eukaryote to have been sequenced with a fairly small genome of a length of 12.1 Mbp (mega base pairs, i.e. nucleotides) encoding approximately 6000 genes [53]. It is unicellular and because it shares homologs, i.e. evolutionarily connected, genes with human, knowledge gained about *Saccharomyces cerevisiae* has value for the study of human genes, in particular diseas involved genes. As of April 2007 the SGD Project [28, 123] and the Gene Ontology Consortium [8, 32] report that out of 5790 ORFs (open reading frames) of the yeast *Saccharomyces cerevisiae* 3698, 4400 and 4946 are annotated to a term of the controlled vocabularies describing the molecular function, the cellular component, or the biological process, respectively [33]. Yet, many annotations are based on large-scale analyses, and caution seems advisable as to their correctness [122]. And

the number of uncharacterised ORFs has recently reduced much more slowly than before. This lack of knowledge about protein function entails considerable uncertainty in large-scale networks of systems biology [56].

For what concerns the type of interactions in the cellular networks, they can only be known by investigating each actual interaction on its own. Enzyme assays examine the kinetics of the reaction an enzyme catalyses measuring either the consumption of substrate or the yield of product over time. Techniques include initial rate measurements or progress curve experiments. When enzyme assays are conducted and documented correctly both the kinetic mechanism and its parameters it are determined [47]. In the literature it happens however that enzyme kinetic parameters are given without any information on the mechanism. The STRENDA Commission endeavours to establish standards for the reporting of enzyme assay results for them to be of optimal use for scientists.

Other interactions like protein complex forming or activation and deactivation of proteins by kinases and phosphatases are often modelled with mass action kinetics.

A given model of a cellular network with fully parametrised interactions should be compared to experimental results. This means that an experiment should measure the concentrations of all involved components, preferably across a time interval of interest. The only one among the concentrations of the constituents of the cellular network we know is that of the genes, there being one genome per cell.

Other components of cellular networks like mRNA, proteins and metabolites need to be measured to get an impression of their abundance under certain environmental conditions. As measurements of abundances at certain time points are snapshots of the momentary state of a cell, they do not as such reveal any dynamic information. Snapshots of the protein total have medical value by discovering so-called biomarkers that indicate a particular disease as well as possibly a susceptibility to a treatment. In Alzheimer's disease plaques of proteins build up in the brain. The protein, amyloid beta protein, can be examined for increased levels with proteomic techniques.

To gain dynamical information, several snapshots taken at intervals, so-called time series, are needed. They can be compared to the outcome of a model simulation. If a model is not fully parametrised one can also use a time series of snapshots for parameter estimation. This is because the path from one snapshot to the other is governed by the system's interactions and thus by the parameters used to describe them.

Current technologies for measuring components of the cellular network include microarrays to measure genome-wide relative mRNA concentrations. Quantitative proteomics based on mass spectrometry yield relative abundances of proteins. Measurements of the entire metabolite content of cells, the so-called metabolome, have to date not yet been accomplished. The major hurdle for metabolomics is to set up an inventory of metabolic compounds. A first attempt to this end has been undertaken by the Human Metabolome Project that has collected information on about 2500 metabolites into the Human Metabolome Database [145].

1.4 Standards for life sciences

> "By the late 1950's it had become evident that the nomenclature of enzymology, in the absence of any guiding authority, in a period when the number of known enzymes was increasing rapidly, was getting out of hand. The naming of enzymes by individual workers had proved far from satisfactory in practice. In many cases the same enzymes became known by several different names, while conversely the same name was sometimes given to different enzymes. Many of the names conveyed little or no idea of the nature of the reactions catalysed, and similar names were sometimes given to enzymes of quite different types. To meet this situation, various attempts to bring order into the general nomenclature of enzymes, or into that of particular groups of enzymes, were made by individuals or small groups of specialists, but none of the resulting nomenclatures met with general approval.
>
> In view of this state of affairs, the General Assembly of the International Union of Biochemistry (IUB) decided, during the third International Congress of Biochemistry in Brussels in August, 1955, to set up an International Commission on Enzymes."

This quotation from the Enzyme Nomenclature 1992 [142] describes a situation where standards have not yet been set up and what confusion ensues. Science as an organised body of knowledge requires conventions of knowledge representation so that this knowledge can be shared and exchanged. This is fundamental to science in general and to bioinformatics and systems biology as young emerging fields of research in particular. Standards for a field of knowledge provide a common language that enables scientists to communicate unambiguously about pieces of knowledge of that field.

The increasing amounts of data being generated in biology by high-throughput techniques offer the chance to gain a view of entire biological systems. These data are available to bioinformaticians and biostatisticians who set up databases and develop comparative tools that categorise and structure data and find connections between data sets. These data serve modellers in systems biology who integrate them into their models of biological networks seeking to simulate entire biological systems.

Today, with increasing amounts of omics data, and with the internet allowing for data exchange at high rates, the need for standards is even more urgent. The challenge is to make sensible use of the data that are produced in volumes requiring electronical means for processing and integration. Data from high-throughput experiments are stored in databases to offer widespread access to them.

Not only standardised ways of talking about a domain of knowledge are needed as ever more knowledge is stored electronically. Electronical storage requires the standards to be translated into electronical formats in order to enable electronical communication between tools generating, storing, analysing and utilising the data. A prominent and widely used electronical format is XML. SBML (Systems Biology Markup Language), a format for encoding dynamical models of biological interaction networks, is a derivative of XML.

Electronical exchange formats themselves are based on object models and ontologies for a specific field of knowledge. These define a vocabulary to represent shared knowledge about objects of a field of knowledge and their relationships; the so-called universe of discourse. The definition provides a nomenclature and system of identification of the objects. ChEBI (Chemical Entities of Biological Interest) [42] is an example of an ontology defining the universe of the so-called small biochemical compounds. The ongoing development of numerous ontologies and the corresponding electronical formats as well as of new standardisation projects have resulted from the need to share knowledge. These ontologies are available from the Open Biomedical Ontologies (OBO) website [35].

The construction of those databases, on the other hand, requires formats of representation of the stored information. These formats correspond to specific object models that conceptualize a domain of research and determine the structure of the stored information. In the area of microarray gene expression data, for instance, the MIAME guidelines [24] for reporting of experimental results have facilitated [26] a flow of data from laboratories into public repositories such as ArrayExpress [25, 10] and Gene Expression Omnibus (GEO) [12, 50].

Information formats of databases contain an inherent data model. For experimental data the formats specify the necessary information for meaningful interpretation of the data. This is necessary if experiments are to be reproducible.

Scientists have to sit down and agree on standards. In the era of systems biology newly developed standards are emerging for purely biological domains, for information on experimental data and for models of biological networks. One of the first initiatives along this lines for experimental data was MIAME which standardised the way of reporting about Microarray experiments [24]. There are several follow-up standardisation projects, e.g. MIAPE [133] or STRENDA [7]. A prominent example of a standardisation of biological knowledge is the GO (Gene Ontology) [8, 32] which provides a controlled vocabulary to assign attributes to genes and gene products in any organism. GO has been widely accepted and is used, for instance, by the SGD project [123].

For dynamical models of biological networks standards have been developed and summed up in the MIRIAM recommendations [100]. MIRIAM [43] is a set of guidelines for annotation of dynamical models with emphasis on the identification of model components such as metabolites or enzymes. Model standardisation is necessary if models are to be comparable, expandable and exchangeable [94]. To this end the Systems Biology Markup Language (SBML) has been developed which is format for the publication of models [74]. A standard for model annotation called MIRIAM has been set up [100].

2 Metabolic Networks

2.1 Metabolism

Metabolism is the set of all biochemical reactions taking place in a cell. The reactions group into pathways by sharing metabolites. The pathways fulfil functions of the living cell allowing it to grow and reproduce by conveying biomaterial to its destination. Generally they are divided into two classes, catabolism and anabolism. Catabolic pathways take up molecules from the environment, so-called fuel molecules like carbohydrates, fatty acids or proteins and break them down into their parts thereby releasing energy in the form of ATP. An example of a catabolic pathway is glycolysis in which the cell gains two molecules of ATP and NADH from one molecule of glucose by being metabolised into two pyruvate molecules. The anabolic pathways, on the other hand, use the components of the macromolecules to construct new molecules that the cell needs, thereby consuming energy. Amino acid synthesis or pyrmidine synthesis are examples of anabolic pathways. The citric acid cycle plays a role in both catabolism and anabolism by burning fuel molecules in aerobic respiration and delivering precursor molecules for amino or fatty acids.

The biochemicals participating in metabolism can be devided into two groups. Those molecules that are consumed and produced in reactions in the course of the metabolic pathways are the metabolites. They constitute the metabolic flux to yield an end product. The other biochemicals take part in the reactions but leave them unchanged or get produced in one and consumed in another reaction. Enzymes, for instance, catalyse the reactions and drive the flux through pathways. Coenzymes are carriers of functional groups required or generated by a reaction. Coenzymes take them up or loose them in the course of a reaction. ATP is a ubiquitous coenzyme being produced in catabolic pathways from ADP and a phosphate group. In anabolic pathways a phosphate group from ATP is consumed leaving ADP.

2.2 Kinetic models of metabolic networks

The cell is an open system. It is embedded in an environment in which constantly changes occur that affect the cell. From this environment it extracts its nutrients. But it is also exposed to stress coming from the environment or receives hormones from other cells. The cell has to respond to these external influences and at the same time maintain its viability. It has to ensure essential functions like cell growth and cell cycle in an environment that constantly changes.

The complex apparatus that the cell uses to communicate with the environment is

the network of signalling pathways. The signalling pathways process and respond to the information the cell receives from its environment. The receptors form the interface. The stimuli are sensed by the receptors and induce the signalling pathways. At the end of a cascade of protein-protein interactions activated transcription factors act upon target genes leading to altered transcription with changed levels of abundances of the corresponding proteins. Among those proteins there are possibly further transcription factors, possibly also enzymes or other proteins. The enzymes will alter the flux through metabolic pathways they control to yield a change in the abundance of the products of the pathway needed by the cell for homeostasis, growth and reproduction.

Metabolism is a dynamical process. Mass constantly flows along the metabolic pathways. In a constant environment metabolic fluxes will not change over time. For instance in one supplying enough nutrient, a population of yeast cells will ideally grow exponentially for some time as each single cell grows at a constant rate. Upon changes in the environment, the cell responds by adapting the fluxes like those producing material for growth during a transient phase and will end up in another steady state if no further environmental changes occur.

Modelling the metabolism of a cell will have to account for its dynamical nature. The fundamental quantities of metabolic models are the concentrations of the biochemical species, basically enzymes and metabolites, and the reaction rates that measure the conversion of biochemicals. In the modelling of chemical reaction systems two approaches have been developed. Both approaches assume that molecules diffuse freely in the regarded compartment and that the number of reactions occurring in a time increment δt is proportional to the number of collisions of the reactants in δt that exceed a certain kinetic energy threshold beyond which the reaction takes place.

The two approaches could be called the microscopic and the macroscopic one. Let us take a look at a simple reaction

$$A + B \xrightarrow{k} X \tag{2.1}$$

with reaction constant k.

Stochastic kinetics

The microscopic, commonly called the stochastic approach counts the molecule numbers of each species and takes into account each single reaction event. Collisions per time interval δt are interpreted in terms of probabilities, i.e. the probability of a reaction occurring is proportional to the number of appropriate collisions:

$$\text{probability of reaction in } \delta t \quad \propto \quad \text{collisions in } \delta t \; .$$

The number of collisions itself is proportional to the number of possible reaction partners, $A \cdot B$ in the case of reaction (2.1) where A and B stand for the molecule numbers of the respective species. For the above reaction (2.1), the probability of a reaction event

occurring within a time increment δt is then [59]:

$$\text{probability of reaction in } \delta t = k \cdot A \cdot B \cdot \delta t + o(\delta t^2) \tag{2.2}$$

with k the reaction constant. If a reaction event takes place the number of molecules will change to $A - 1$ and $B - 1$.

Deterministic kinetics

The macroscopic, commonly called the deterministic approach does not account for every single reaction event and usually measures amounts in concentrations. It interprets the collision frequencies in terms of rates, i.e.

$$\text{rate of reaction at } t \quad \propto \quad \text{collisions at } t \ .$$

It is viewed as a limiting case of the stochastic approach for infinite molecule numbers. Fluctuations and correlations between species are disregarded. Each reaction is assigned a rate. For the above reaction (2.1), the rate, similarly to the stochastic approach, then is

$$\text{rate of reaction } \ = \ k \cdot V \cdot A \cdot B \tag{2.3}$$

where V is the volume of the system and k the reaction constant. It is included because in the deterministic approach the variables A and B stand for the concentrations of the respective species. If we denote the reaction rate by v, then the concentrations of species A and B will change by $dA = dB = -v\ dt$ in a time increment dt.

2.2.1 Deterministic modelling

In deterministic modelling, the temporal evolution of metabolic networks is described by a differential equation system for the state S which is the vector of concentrations of the biochemical species. The concentration of each species changes with the reactions it participates in. This change is a linear combination, a weighted sum over the respective reaction rates where the weights are the stoichiometric coefficients of the species in those reactions. Thus, the temporal evolution of the state S can be expressed as the product of a matrix and a vector:

$$\frac{d\mathbf{S}}{dt} = N \cdot \ v(\mathbf{S}, p) \ . \tag{2.4}$$

N is the stoichiometric matrix and encodes the structure of the metabolic network. The rows denote the species and the columns the reactions. The entries in a row shows the reaction a species is involved in either as a reactant or as a product. The entries in a column, on the other hand, say which species are consumed or produced in a reaction and how many molecules of each are involved. The reaction rates are denoted by v and depend on the concentrations of the species S and parameters p, e.g. reaction constants for the case of mass action and intial concentrations of the biochemical species.

2.3 Time scales and dimensional reduction

Metabolism involves a vast span of time scales across several orders of magnitude (s. Tbl. 2.1). The fastest processes include binding of substrates to enzymes that happen on the micro- to millisecond time scale [14, 44, 47]. Once the substrate is bound to the enzyme it is turned over into product within a fraction of to tens of milliseconds [14, 44, 47]. Metabolic pathways reach their steady state on the scale of tens of seconds up to an hour [93, 136]. Regulation through enzyme synthesis in response to environmental changes takes tens of minutes up to several hours [39, 137, 29]. Protein degradation acts on a wide range of time scales encompassing long-lived proteins that last up to 10^6s down to short-lived ones with a life-time of seconds.

The main players in metabolism, the enzymes and metabolites, undergo temporal change on different time scales. Models of metabolic pathways focus on the dynamics of the concentrations of metabolites and the distribution of metabolic fluxes in steady state. The enzymes, on the other hand, are often taken not to or to change much more slowly in their concentrations.

By a change in concentration, enzymes exert control on the fluxes through them. The biotechnological industry, for instance, is interested in those enzymes that exert the major control on a pathway in order to engineer strains with improved antibiotics yields while at the same time avoiding by-product formation.

The main processes in metabolism are enzyme-substrate binding, substrate turnover and change of enzyme conentration levels. The processes can be categorised into three groups according to the time scales on which they happen.

1. The fast processes compared to metabolite turnover rates are substrate-enzyme binding processes.

2. Metabolite turnover is the process of interest in models of metabolic networks. Thus, its time scale is in the focus of any model of a metabolitc network.

   For instance in the yeast *Saccharomyces cerevisiae*, with a cell volume of approximately $70\mu\text{m}^3 = 0.07\text{pL}$ [107], a substrate concentration of 1mM [2] would correspond to $4 \cdot 10^7$ molecules in the cell. With an enzyme taking 10^{-3}s to turn over a substrate molecule and assuming an enzyme concentration of 1μM ($4 \cdot 10^4$ enzymes per cell), it would take approximately 10s to convert the amount of substrate into product, thereby neglecting reversibility of the process and incomplete enzyme saturation.

3. Slow processes relative to metabolite turnover are gene expression and enzyme synthesis. Protein degradation is disregarded in models of metabolism. For metabolism in the yeast *Saccharomyces cerevisiae*, for instance, enzymes belong to the stable proteins [13].

Different time scales in a modelled system entail mathematical simplicifications. It is common to introduce a scaling parameter ϵ with $0 < \epsilon \ll 1$ [46, 30] to separate the time

Time scales in metabolism

process	time scale (s)	reference
substrate-enzyme binding	$< 10^{-4}$	[14, 44, 47]
enzyme turnover	10^{-4}–10^{-2}	[14, 44, 47]
pathway transition to steady state	$10 - 10^{3}$	[93, 136]
gene expression	10^{3}–10^{5}	[39, 137, 29]
protein degradation	1–10^{6}	[139, 13]

Table 2.1: Processes in metabolism occur on a wide range of time scales.

scales. If the time scale of interest is denoted by t then the time scale of fast processes becomes

$$t_f = \frac{t}{\epsilon} \tag{2.5}$$

and the time scale of slow processes

$$t_s = \epsilon \cdot t \; . \tag{2.6}$$

Accordingly, we separate the dynamical system variables into fast variables x_f, the variables of interest x, and the slow variables x_s. If on their respective time scales the variables change according to

$$\begin{aligned} \frac{dx_f}{dt_f} &= g_f(x_f, x, x_s, p) \\ \frac{dx}{dt} &= g(x_f, x, x_s, p) \\ \frac{dx_s}{dt_s} &= g_s(x_f, x, x_s, p) \end{aligned}$$

then, on the time scale of interest t, the equations become

$$\frac{dx_f}{dt} = \frac{1}{\epsilon} \cdot g_f(x_f, x, x_s, p) \tag{2.7}$$

$$\frac{dx}{dt} = g(x_f, x, x_s, p) \tag{2.8}$$

$$\frac{dx_s}{dt} = \epsilon \cdot g_s(x_f, x, x_s, p) \tag{2.9}$$

$$\tag{2.10}$$

As $0 < \epsilon \ll 1$, we can immediately see that the slow variables x_s hardly change, that is $dx_s/dt \approx 0$, or

$$x_s(t) \approx x_s^0 = \text{const} \; . \tag{2.11}$$

On the other hand, the fast variables change very rapidly as $1/\epsilon$ will be very big and

change until $g_f(x_f, x, x_s) \ll \epsilon$ or

$$g_f(x_f, x, x_s) \approx 0 \, . \tag{2.12}$$

This last condition, Eq.(2.12), imposes algebraic constraints on the solution to the differential equation Eq.(2.8) and therefore on the dynamics of the variable of interest x.

Michaelis-Menten model

In a famous example of separation of time scales the Michaelis-Menten kinetics were derived for an irreversible enzyme-catalysed reaction involving one substrate and one product:

$$S + E \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} P \tag{2.13}$$

When this system is modelled according to Eq.(2.4) with reaction velocities described by mass action kinetics, it comprises the four dynamic components S, E, ES, P, with the three reaction rates k_{+1}, k_{-1} and k_2 and four initial values of the components as parameters. The parameters k_{+1} and k_{-1} describe the rate of enzyme-substrate association and dissociation, respectively. The parameter k_2 is called the turnover rate of the enzyme. The four dynamical variables reduce to two because of the conservation relations $E_T = E + ES = \text{const}$ and $S + P = \text{const}$.

Basis for the time scale separation in this system is the assumption the reversible binding of the substrate to the enzyme occurs at a much faster rate, described by the rate constants k_{+1} and k_{-1}, than the turnover of the bound substrate and subsequent release of product, described by the rate constant k_2. Thus, in modelling this conversion of substrate S to product P, the following simplification is made. Because the reversible process

$$S + E \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} ES \tag{2.14}$$

is much faster than the irreversible process

$$ES \overset{k_2}{\rightarrow} E + P \, , \tag{2.15}$$

it is assumed that reaction (2.14) reaches a steady state before reaction (2.15) has had a noteworthy effect on the system dynamics. This is called the quasi-steady state assumption.

Schematically, after time scale separation the reaction can be represented in the sim-

Example of reduction of model dimension

kinetics	species	conserved quantities	parameters	dimension
MA	S, E, ES, P	$S+P$	k_{+1}, k_{-1}, k_2	2
		$E_T = E + ES$		
MM	S, P	$S+P$	E_T, k_2, K_m	1

Table 2.2: Processes in cellular biology span a vast range of time scales. When systems are modelled that encompass very different time scales, simplifications in mathematical description can be made through time scale separations. These reduce the size of the phase space and the computational cost of simulation. In a classical example, time scale separation was used in the description of the irreversible enzyme catalysed reaction $E + S \rightleftharpoons ES \longrightarrow P$ that reduces the model dimension from 2 to 1. This lead to the famous Michaelis-Menten kinetics. (Abbreviations are MA: mass action, MM: Michaelis–Menten.)

pler form

$$S \xrightarrow{(E_T, K_m, k_2)} P \,. \tag{2.16}$$

For the mathematical formula of the reaction kinetics see Eqs.(2.18), for its derivation App.6.

The model corresponding to scheme (2.16) contains two dynamical variables S and P, constrained by the conservation $S + P = \text{const}$, and following parameters: two kinetic constants K_M and k_2, the total enzyme concentration E_T and two initial values for S and P. The new parameter K_M is called the Michaelis-Menten constant. The ratio of the Michaelis-Menten constant K_m and the substrate concentration S determine the enzyme saturation (s. Eq.(2.20) and explanation thereafter).

In this description of a biochemical reaction in a metabolic pathway, the enzyme is not a modelled species anymore. Rather its total concentration has become a parameter of the reaction kinetics. Time scale separation reduces the dimensionality of a modelled system (s. Tbl. 2.2) and thus saves computational costs.

For a derivation of the kinetics of this irreversible enzyme-catalysed reaction through time scale separation see appendix 6.

2.4 Kinetics

Metabolic networks can be considered from two viewpoints. On the one hand, they can be regarded as consisting of enzymes that interact by converting and passing on metabolites from one to the next. On the other hand, the metabolites can be taken as the nodes of the metabolic network that interact through biochemical reactions catalysed by enzymes.

Deterministic modelling described by Eq.(2.4) in general is a combination of these views. The nodes can be any biochemical molecule. The interaction can be transport, binding, conversion or activation processes, subsumed under the notion of reaction. The momentary concentrations of the biochemical molecules, encoded by the state S, change through the reactions. The changes brought about by the reactions are expressed by

the reaction velocities $v(S, p)$. They depend on the state S itself and on the type of interaction encoded in the functional form of v and its parameters p.

In a crude picture of catalysis, the metabolites bind to the enzymes which leads to a conformational change of the enzyme facilitating the actual biochemical reaction. Each enzyme mechanism is a special case of this description of catalysis, often involving further details. Reversibility can come into play. Metabolites of the same species can cooperate. Effectors can influence the binding properties of an enzyme, can strengthen, weaken or even prevent binding.

Once the biomolecules bound to the enzyme are close enough for a reaction to happen, series of successive events occur on different time scales and submolecular levels. The modelling of biochemical reaction networks focusses on changes of the concentrations of biochemical species and not on the details of how the reactions actually occur. It is thus of importance to know the involved reactants, any details on submolecular levels are disregarded and lumped into single rates. In fact, assigning a kinetic rate to a process is a phenomenological black box description of the process.

The simplest form of kinetics in deterministic modelling of metabolic networks is mass action kinetics. Mass action assigns a rate to each interaction between molecules that is dependent on the concentrations of the reactants and the stoichiometry of the process. Equilibria of reversible processes are characterised by the ratio of the rates in the two directions. For instance, a reversible process between two states A and B going at rates k_+A and k_-B,

$$A \underset{k_-}{\overset{k_+}{\rightleftharpoons}} B \ ,$$

has an equilibrium characterised by $B_{eq}/A_{eq} = k_+/k_-$. This is called the law of mass action [41].

In principle, any chemical reaction system can be modelled with mass action kinetics. As cellular processes involve many different time scales (s. Sec. 2.3), modelling metabolism has to account for the processes happening on these time scales. But modelling all processes involved, from slow to fast, with mass action kinetics becomes cumbersome and computationally costly. Describing the dynamic evolution of every single chemical species of the reaction system, including the intermediate binding complexes of metabolites and enzymes, blows up the dimension of the differential equation system. And the numerical effort increases not only with the dimension of the model, but also by integrating across different time scales.

To dodge these difficulties in the modelling of metabolic networks, fast binding of substrates and effectors compared to the conversion of substrate into product is assumed. Furthermore, effects on enzyme concentrations due to gene expression are often disregarded on the time scale of the catalysed biochemical reactions [30].

The simplest case of an enzyme-catalysed reaction is an irreversible reaction turning a single substrate molecule A into a product molecule B:

$$S \longrightarrow P \ . \tag{2.17}$$

After separation of time scales (s. Secs. 2.3 and 6), the kinetics of this reaction are described by the Michaelis-Menten equation:

$$\frac{dP}{dt} = k_2\, E_T\, \frac{S}{K_m + S} \tag{2.18}$$

$$= v_{max}\, \frac{S}{K_m + S} \tag{2.19}$$

$$= v_{max}\, \frac{\tilde{S}}{1 + \tilde{S}}\,, \tag{2.20}$$

with v_{max} the maximal velocity of the reaction for a given total enzyme concentration, K_m the Michaelis-Menten constant and $\tilde{S} = S/K_m$.

The last version of Michaelis-Menten kinetics, Eq.(2.20), rescales the substrate concentration S in terms of the Michaelis-Menten constant. It emphasises that it is the ratio $\tilde{S} = S/K_m$ which determines the saturation of the enzyme

$$\frac{ES}{E_T} = \frac{S/K_m}{1 + S/K_m}\,.$$

Irreversibility of processes does not generally hold, but can be used as an approximation as long as the substrate concentrations exceed those of the products by far. Actually, more complex mechanisms are mostly at work accounting for more than one molecule binding to an enzyme. The above reaction (2.17) in the reversible case,

$$S \rightleftharpoons P\,,$$

can be modelled by the following phenomenological kinetics [92]

$$v = \frac{v^+_{max}\,(S/K_{mS}) - v^-_{max}\,(P/K_{mP})}{1 + S/K_{mS} + P/K_{mP}}\,. \tag{2.21}$$

The large majority of enzymes actually catalyses reactions with two substrates [47]. When more than one substrate come into play several questions arise. The enzyme can require a certain order of substrate binding. The enzyme Aspartate transcarbamylase for instance, catalysing the reaction

$$\text{carbamoyl phosphate} + \text{aspartate} \rightleftharpoons \text{carbamoyl aspartate} + \text{phosphate}\,,$$

has a compulsory-order kinetic mechanism with Carbamoyl phosphate binding first [47]. Other enzymes require the second substrate to bind only after the first has been released where the enzyme is left with a chemical entity that the second substrate subsequently takes up. This is called the double-displacement mechanism. One speaks of a random-order mechanism if any temporal order of the binding of substrates is possible for catalysis.

These kinetics can become intricate when real metabolic pathways. Teusink et al. [136] model the glycolytic pathway. One reaction of the pathway catalysed by the enzyme

Fructose-bisphosphate aldolase (EC 4.1.2.13) is

$$\text{F1,6BP} \rightleftharpoons \text{GADP} + \text{DHAP} \tag{2.22}$$

where F1,6BP denotes D-Fructose 1,6-bisphosphate, DHAP stands for Dihydroxyacetone phosphate and GADP means D-Glyceraldehyde 3-phosphate. The catalysed reaction is assumed to follow an ordered uni-bi mechanism where the substrate F1,6BP splits into the products DHAP and GADP with GADP being released first. If we abbreviate the reaction (2.22) by $S \rightleftharpoons P + Q$, where the substrate is denoted by S and the products DHAP by Q and GADP by P, respectively, the detailed reaction scheme reads

$$\text{E} \rightleftharpoons \text{ES} \rightleftharpoons \text{EPQ} \rightleftharpoons \text{EQ} \rightleftharpoons \text{E}$$

with kinetics [136, 101]

$$v = E_T \, \frac{k_{cat}^{+} \, \frac{S}{K_S} \, (1 - \frac{\Gamma}{K_{eq}})}{1 + \frac{S}{K_S} + \frac{P}{K_P} + \frac{Q}{K_Q} + \frac{SP}{K_S K_{iP}} + \frac{PQ}{K_P K_Q}} \, . \tag{2.23}$$

The dissociation constants for the respective metabolites are K_S, K_P, K_Q and K_{iP} is called the substrate inhibition constant of substrate P.

There are more aspects to enzyme kinetics. Modifiers can inhibit or activate an enzyme. If the enzyme has multiple binding sites the sites can have an influence on each other's binding properties leading to cooperativity and Hill kinetics (s. e.g. [47]).

A random-order kinetic mechanism is the so-called convenience kinetics [103]. The kinetics reads

$$v = E_T \cdot \left(\prod_A \frac{\bar{c}_A}{\bar{c}_A + 1} \right) \cdot \left(\prod_I \frac{1}{\bar{c}_I + 1} \right) \cdot \frac{k_{cat}^{+} \prod_S \bar{c}_S^{\,n_S} - k_{cat}^{-} \prod_P \bar{c}_P^{\,n_P}}{\prod_s \left(\sum_{m=0}^{n_S} \bar{c}_S^{\,m} \right) + \prod_p \left(\sum_{m=0}^{n_P} \bar{c}_P^{\,m} \right) - 1} \, . \tag{2.24}$$

and is defined for biochemical reactions with any number of substrates and products. The index variables A, I, S and P run over the sets of activators, inhibitors, substrates and products of the reaction, respectively. The total concentration of the enzyme catalysing the reaction is denoted by E_T. The variables k_{cat}^{+} and k_{cat}^{-} stand for the turnover rates of the enzyme in the forward (+) and the backward (−) direction, respectively. The rescaled concentrations of substrates and products, $\bar{c}_S = c_S / k_{mS}$ and $\bar{c}_P = c_P / k_{mP}$, are the ratios of the substrate and product concentrations with their K_m values, and for activators and inhibitors, $\bar{c}_A = c_A / K_{aA}$ and $\bar{c}_I = c_I / K_{iI}$, the ratios of the activator and inhibitor concentrations with their respective dissociation constant, K_{aA} or K_{iI}. Finally, n_S and n_P are the stoichiometric coefficients of substrate S and product P, respectively.

This formula is directly applicable once the stoichiometry of the enzyme-catalysed reaction and the numbers of activators and inhibitors are known. Effectors enter the

kinetic formula in Eq. (2.24) as factors only.

The number of kinetic parameters that enter this kinetic expression are one K_m value for each reactant, one K_a value for each activator of the enzyme, a K_i value for each inhibitor and two turnover rates $k_{cat}^{\pm}$ for the enzyme. The total enzyme concentration E_T also needs to be known.

The velocity of reaction (2.22) expressed with convenience kinetics reads

$$v = E_T \, \frac{k_{cat}^{+} \, \frac{S}{K_S} \; - \; k_{cat}^{-} \, \frac{PQ}{K_P K_Q}}{1 + \frac{S}{K_S} + \frac{P}{K_P} + \frac{Q}{K_Q} + \frac{PQ}{K_P K_Q}} \; . \tag{2.25}$$

2.5 Thermodynamics and Kinetics

Living organisms are open systems. They exchange both matter and energy with their environment. So laws from classical thermodynamics that concern closed systems do not apply in the case of living organisms. In fact they maintain their state of high order, i.e. low entropy, by increasing the entropy of their environment. But equilibrium thermodynamics still has some consequences for non-equilibrium living systems.

In following the concentrations of species S_i will be denoted by c_i contrary to the usual habit in metabolic modelling as S denotes the entropy in thermodynamics.

When considering biochemical processes the Gibbs free energy G is the appropriate energy as it is a function of the temperature T, the pressure p and the concentration c_i in mols of the species S_i the system is composed of:

$$G = G(T, p, c_i) \; .$$

If the system variables T, p or c_i are modified the Gibbs free energy changes according to

$$dG = -S dT + V dp + \sum_i \mu_i dc_i$$

where S is the entropy of the system and the μ_i are the chemical potentials of the species S_i.

For a biochemical reaction in equilibrium at constant temperature T and constant pressure p the condition

$$0 \stackrel{!}{=} \Delta G \; = \; \frac{\partial G}{\partial \xi} \; = \; \sum_i n_i \mu_i \; . \tag{2.26}$$

holds, where $dc_i = n_i d\xi$ and ξ the progress variable or extent of reaction which is measured in mols. Now for an infinite dilution, i.e. where the activity numerically equals the concentration, the chemical potential can be written as [72]

$$\mu_i - \mu_i^0 = RT \ln[c_i] \tag{2.27}$$

where $[c_i]$ is the dimensionless activity on the molarity scale and numerically equal to

the concentration c_i in mols of species S_i. In what follows we will write c_i instead of $[c_i]$. The quantities μ_i^0 are called the standard Gibbs free energies of species S_i.

Single reaction

If Eq. (2.27) is substituted into Eq. (2.26) this results in

$$0 = \Delta G \;=\; \Delta G^0 + RT\ln\left(\prod_i \left(c_{i,eq}\right)^{n_i}\right) \tag{2.28}$$

where

$$\Delta G^0 = \sum_i n_i \mu_i^0 \tag{2.29}$$

and $c_{i,eq}$ denotes the concentration in mols of species S_i at equilibrium. Eq. (2.28) is equivalent to

$$\Delta G^0 = -RT\ln\left(\prod_i \left(c_{i,eq}\right)^{n_i}\right) \stackrel{!}{=} -RT\ln K_{eq} \tag{2.30}$$

where K_{eq} is the equilibrium constant. At equilibrium, the net rate of a reaction vanishes. $\Delta G = 0$ means that

$$v(c_{i,eq}, p) = 0\ . \tag{2.31}$$

According to Eq. (2.31) equilibrium thermodynamics imposes an algebraical constraint on the kinetic parameters p. The equilibrium concentrations $c_{i,eq}$ are constrained by the equilibrium constant in Eq. (2.30). So in general, this will also put constraints on the parameters p such that Eq. (2.31) is fulfilled.

Reaction network

In the case of a metabolic network with coupled reactions $j = 1...N$ the dependencies of the kinetic parameter spread over the entire network. For each reaction j at equilibrium

$$\Delta G_j^0 = -RT\ln\left(\prod_i \left(c_{i,eq}\right)^{n_{ij}}\right) = -RT\ln K_{eq,j} \tag{2.32}$$

and

$$v_j(c_{i,eq}, p_j) = 0\ . \tag{2.33}$$

The equilibrium constant K_{eq} in Eq. (2.32) imposes a constraint on the equilibrium concentrations of the reaction species $c_{i,eq}$ involved in reaction j of the metabolic network. By Eq. (2.33) the constraint on the equilibrium concentrations transmits to the parameters p_j of reaction j. As the reactions of the metabolic network are coupled and some biochemical species take part in more than one reaction, i.e. a row i of the stoichiometric matrix $N = (n_{ij})$ has an entry unequal to zero in at least two columns j_1 and j_2, by Eq. (2.32) the dependence of the kinetic parameters expands across the entire reaction network.

In the field of enzyme kinetics, the equations relating the equilibrium constant of an enzyme-catalysed reaction and the kinetic parameters of the reaction mechanism are called Haldane relationships. They depend on the reaction mechanism and therefore on the model of the reaction. Most Haldane relationships take the general form [126]

$$K_{eq} = \frac{\left(v^+_{max}\right)^m K_{P_1} K_{P_2} K_{P_3} ...}{\left(v^-_{max}\right)^m K_{S_1} K_{S_2} K_{S_3} ...} \tag{2.34}$$

where the substrates are denoted by S_1,S_2,..., the products by P_1,P_2,... The parameters K_{S_1},..., K_{P_1},... are either Michaelis-Menten or dissociation constants and the maximal velocities in the forward or backward direction are named v^+_{max} and v^-_{max}, respectively. The values of the exponent m belong to the set $-1, 0, 1, 2, ...$

Intuitively, the interpretation of the Haldane relationships is that the different processes in the forward and reverse direction involved in enzyme catalysis between the enzyme, i.e. the binding of substrate to enzyme, the chemical turnover by the enzyme, the product release, or binding of an inhibitor, have to occur at rates such that at chemical equilibrium the overall rates of the biochemical reaction in the forward and reverse directions cancel each other out and the equilibrium concentrations satisfy the law of mass action.

Haldane relationships are useful for checking the consistency of kinetic parameters or even for accepting or ruling out certain kinetic mechanisms when examining a certain enzyme [126].

As we will discuss later (s. Ch. 5) they are not only helpful in enzyme assays when kinetic parameters are measured for an enzyme catalysing a certain reaction. They also impose constraints in kinetic modelling. Kinetic models of moderate size contain tens of parameters and require a great effort to be parametrised. In most cases literature data cannot not be found so that parameter estimation techniques have to serve. Haldane relationships mean that not any parameter combinations can be chosen for the parameters of a network because they have to fulfil the the Haldane relationship of each reaction in the network.

3 Kinetics require parameters

In Sect. 2.4 the phenomenological description of molecular interactions was introduced. The simplest description of interactions is mass action kinetics. Often kinetics derived from mass action, based on the assumption of fast enzyme-binding are employed. Michaelis-Menten kinetics is the simplest example of such a derived enzyme mechanism. Others kinetics have been developed describing more complex enzyme catalysis.

Any kinetic description of cellular processes requires parameters. When processes are described by mass action a reaction constant k is assigned to each interaction accounted for in the model (s. Eqs. (2.2) and (2.3)). The Michaelis-Menten description of an irreversible enzyme-catalysed reaction requires, according to Eq.(2.18), as parameters the enzyme concentration E_T, the turnover rate of the enzyme k_2 and the Michaelis-Menten constant K_m.

When a model is constructed, the parameters need to be assigned numerical values so that the model can serve its purpose. Its dynamics and steady state should be compared to the biological system and, preferably, novel properties of the system are the outcome of the modelling that then can be tested experimentally.

The number of parameters in models of metabolic pathways is often quite considerable. Moderate size models require tens of paramters. A model of glycolysis by Rizzi et al. with 20 reactions uses more than 80 kinetic parameters [128]. In another model of glycolysis, Teusink et al. include 19 reactions, thereof two at equilibrium and two further described by constant rates, requiring nearly 60 kinetic parameters [136].

So how are the parameters assigned numerical values such that the kinetics describe the cellular processes as well as possible? There are three ways of gaining parameter values. First, the parameters can be measured together with the identification of the enzyme mechanism. Second, a literature search for enzymes and their kinetics can yield information on the kinetic mchanism and its parameters. Third, parameter estimation techniques for the kinetic parameters are employed when in lack of data.

A classical approach to measure enzyme kinetic mechanisms and their parameters are enzyme assays [105]. In enzyme assays either yield of product or consumption of substrate are measured as a function of time. The measurements are repeated at several substrate concentrations that are preferably not very large as compared to the K_m value of the substrate in order to avoid saturation of the enzyme. Furthermore so-called initial velocity conditions are often chosen to maintain the enzyme-substrate complex at equilibrium. Attention has to be paid to the correct composition of the buffer, its salinity, its pH, cofactors and effectors. For instance, a K_m value which seems unphysiologically high could have been measured in the absence of an activator of the enzyme present *in vivo*. Enzyme assays clarify the kinetic mechanism. And the times series of consumed substrate or yielded product are used in a non-linear regression done with computers to

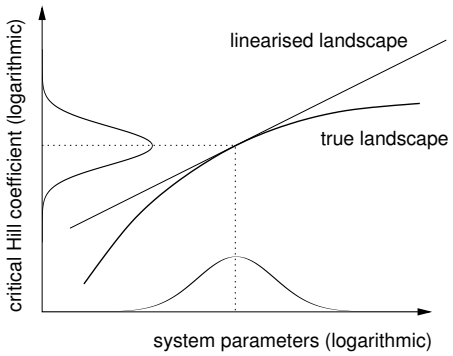

Figure 3.1: The critical Hill coefficient depends on the parameters α, β, and k. The diagram schematically shows $h_{\text{crit}}(\alpha, \beta, k)$ as a function of the parameters. The parameters α, β, and k, represented by the abscissa, are drawn from a random distribtution. Here the distribution is Gaussian on the logarithmic scale (i.e. log-normal distribution of parameters). By computing the value of h_{crit} for each instance of the parameters drawn from the distribution we can sample its distribution (plotted on the ordinate). If $h_{\text{crit}}(\alpha, \beta, k)$ were a linear function of the parameters its distribution would also be a log-normal distribution. This Fig. is from [21]

determine the parameter values. Before computers were common in science, linearisations like the Lineweaver-Burke plot [47] were employed. But they distort error bars as well as the intervall between time points which can heavily distort the outcome of the linear regression.

The second possibility for gaining information on enzyme mechanism and the numerical values of their parameters is literature search. There are also electronical resources like BRENDA [134] that collect enzyme kinetic data from literature.

Nevertheless, for many enzymes neither the kinetic mechanism nor the parameters have been characterised experimentally. In this case one has to assume a mechanism and only estimation techniques can deliver values for kinetic parameters. Assigning parameters to non-linear kinetic descriptions of network interactions necessarily requires information that encodes the dynamics of the network. Estimation of kinetic parameters thus is often based on time series data. For enzyme kinetic parameters for instance, time series of metabolites are useful. The model output for a certain parameter input is compared to time series data. In a maximum likelihood estimation, the parameter set that yields the best approximation to the time series data in an iterative search through the parameter space will be chosen as the parameters that describe the data best.

3.1 Structure, parameters and dynamics

Mathematical models of dynamic biochemical processes,

$$\frac{dS}{dt} = N \cdot v(S, p) \ ,$$

are characterised by the network structure N, the kinetics v and their parameters p. As discussed earlier on, there are three possible ways of obtaining a parameters p for a certain interaction: direct measurement, literature search or parameter estimation. In any case, the parameters are uncertain. Experimental data are subject to measurement errors. Parameter estimation results, apart from being based on experimental time series, yields results that depend on the method of scanning the parameter space as well as on the chosen distance measure. To what extent does dynamic uncertainty of models arise from the uncertainty in the parameters?

Given the structure N of a network and its kinetics v the dynamical properties can depend crucially on the values of the parameters. It is of interest to ask how sensitive the dynamical properties of a network model are with respect to uncertainties in the parameters. Or stated differently, how robust are the network properties with respect to changes in the parameters.

For instance, switches that are modelled by sigmoidal signal-response curves, e.g. Hill curves, have been examined for robustness [19]. Switches can be ultrasensitive meaning they exhibit a sharp, step-like on-off-profile dependent on the signal strength. A slight change in signal can make the response switch between two states. Or they can respond more gradually, the response being growing steadily with stronger signal. In the case of the Goldbeter-Koshland switch the sensitivity depends on the level of saturation of the converter enzymes [65], i.e. the ratio of the protein substrate and the K_m describing its binding to the converter enzyme. An ultrasensitive switch is said to robust if its on-off-profile is maintained upon considerable changes in the parameters of the interactions of the module. And more generally, a dynamical property of a network is said to be robust when it is stable with respect to change in the parameters.

Parameters of interactions can change due to mutations of the genes. As kinetic parameters describe binding and conformational properties of proteins they depend on the protein structure and can change upon mutations. Moreover, within a population parameters are distributed due to the genetical heterogeneity of the population. It is thus of interest to ask how changes in parameters influence the dynamical properties of a given network. Goldbeter-Koshland switches and MAPK cascades have been examined in this respect [19].

Dynamical properties of a network depend on the structure of the model, on the interactions and their kinetic parameters [88]. These properties, e.g. the period of a model of circadian rhythms, generally cannot be calculated analytically. Consequently, also the effect of changes in the parameters on the network properties, described by the so-called sensitivities, cannot be assessed analytically.

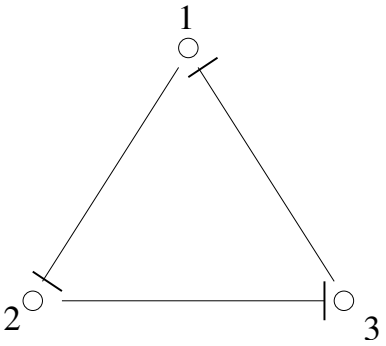

Figure 3.2: The repressilator is a gene regulatory network consisting of three genes inhibiting one another such that the genes form a closed circuit. The network is modelled by Eq.(3.1) with a first-order degradation term and a production term dependent on the presence of inhibitor concentration. The binding of the inhibitor to the gene is described by a Hill curve, i.e. with increasing hill coefficient the inhibition exhibites a more pronounced on/off profile dependent on the inhibitor concentration. This Fig. is from [21]

A way around this problem is drawing parameters from assumed distributions and calculating the dynamical property. This way we can probe the distribution for the respective property. These calculations can be interpreted as mimicking gene mutations that affect parameter values or as representing the genetic spread in a population.

Here we numerically calculate a property a of network for instances of parameters drawn from a distribution and thus find the distribution of the property itself. If the distribution of the property is narrow, we conclude that this quantity is strongly determined by the network structure and less so by the parameters itself, at least for the ensemble of parameters considered. This makes it possible to study which kind of quantitative and qualitative behaviour is to be expected from the model.

Monte Carlo simulations with random parameters have been used to compute the distributions of metabolic concentrations, metabolic fluxes, control coefficients, and other variables [91, 1]. The same approach has been applied to gene regulatory circuits [90] and a MAP kinase cascade [19].

3.2 The repressilator

Here we focus on the bifurcation point of a network model. We study a simple genetic network that has been investigated by Elowitz and Leibler [45] called the repressilator (s. Fig. 3.2). It shows a parameter-dependent transition from a stable steady state to a limit cycle with persistent oscillations, known as a Hopf bifurcation. The Hill coefficient in the kinetic equations is the critical parameter. By sampling all other parameters α,

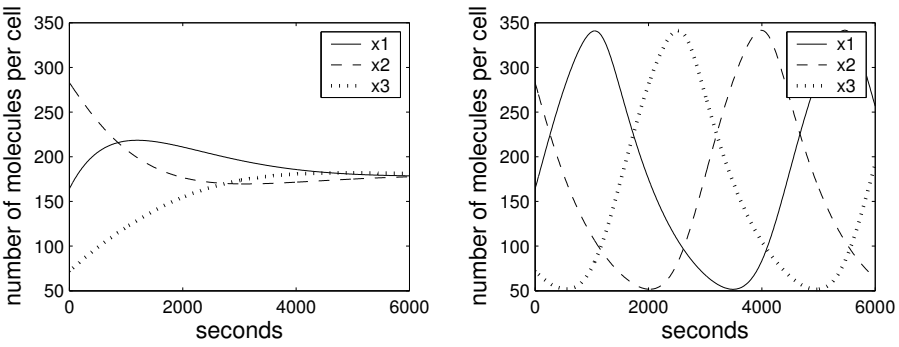

Figure 3.3: Dynamics of the repressilator below and above the Hopf bifurcation. The Hill coefficient h is a bifurcation parameter. For $h = 1 < h_{\text{crit}}$, the system always tends to a steady state according to Eq.(3.2), while for $h > h_{\text{crit}}$, oscillations can arise. The values of the parameters are: $\alpha = 0.001\text{s}^{-1}$, $\beta = 0.5\text{s}^{-1}$, $k = 100$. The critical Hill coefficient is $h_{\text{crit}} \approx 2.78$. Concentrations are given in molecules per cell. This Fig. is from [21]

β and k from predefined distributions, we compute the distribution of the critical value of the Hill coefficient.

The repessilator is a genetic network composed of three genes that form a negative-feedback loop, each gene repressing the transcription of the gene it acts on. Whereas Elowitz and Leibler model the system with six dynamical variables, that is three mRNAs and three proteins, we reduce the model two three variables, the abundances of the gene products, x_i, for genes $i = 1, 2, 3$.

We further simplify the model by making it symmetric and set the parameters expressing the interactions equal for each x_i. The dynamics are thus modelled by the differential equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = -\alpha x_i + \frac{\beta}{1 + (x_{j_i}/k)^h} \quad \text{for } i = 1, 2, 3\,, \tag{3.1}$$

with $h > 0$ describing repression and $j_1 = 3, j_2 = 1, j_3 = 2$ indicating the inhibiting gene for genes $i = 1, 2, 3$. The parameter α describes the rate of degradation and β the full, unhindered transcription. The effect of the genes on each other is described by a Hill signal response-function with dissociation constant k and the Hill coefficient h.

For the symmetric repressilator, modelled according to Eq.(3.1) with $h = 1$, the steady state can be calculated analytically. It reads

$$x_i = \sqrt{\frac{k\beta}{\alpha} + \frac{k^2}{4}} - \frac{k}{2}\,. \tag{3.2}$$

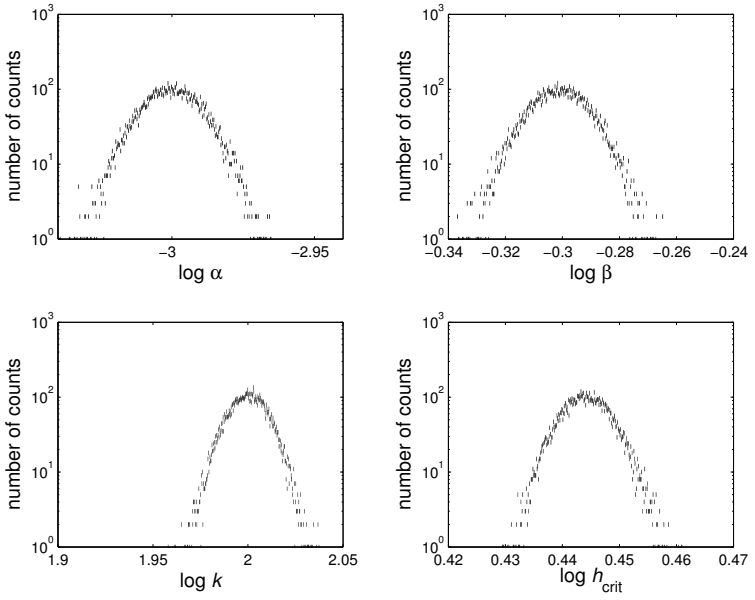

Figure 3.4: Three of the four plots show histograms of the distributions the parameters α, β and k were drawn from. The mean values correspond to $\bar{\alpha} = 0.001\text{s}^{-1}$, $\bar{\beta} = 0.5\text{s}^{-1}$, and $\bar{k} = 100$. The width was in each case $\sigma = 0.01$. The bottom right plot shows the histogram for the resulting distribution of the value h_{crit}. This Fig. is from [21]

The genetic interaction described by a Hill curve

$$1/(1 + (x_i/k)^h) \tag{3.3}$$

describes a repressor for $h > 0$ and an activator for $h < 0$. For $|h| = 1$, it turns into the normal Michaelis-Menten curve. For $|h| > 1$, the signal-response curve becomes a sigmoid with inflection point

$$k \sqrt[h]{\frac{h-1}{h+1}}$$

lying between 0 and k. As the Hill signal-response curve (3.3) takes values between 0 and 1 and is strictly monotonic, we measure the width of the switch by the length of the interval in which the curve falls from 0.9 to 0.1. It is calculated to be

$$x_{0.1} - x_{0.9} = k \left| \sqrt[h]{\frac{0.9}{0.1}} - \sqrt[h]{\frac{0.1}{0.9}} \right| .$$

The width is proportional to k and declines with h. For constant k, the signal-response becomes more switch-like with larger h.

Given values for α, β, and k, this system can undergo a Hopf bifurcation at a certain value $h = h_{\text{crit}}(\alpha, \beta, k)$: for values $h < h_{\text{crit}}$, the system reaches a stable steady state,

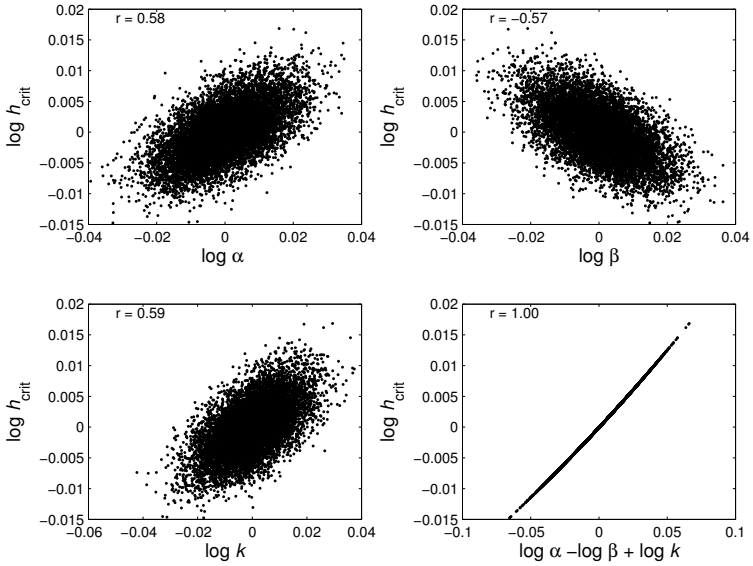

Figure 3.5: Correlation plots for the three parameters α, β and k versus h_{crit}. The log-normal distributions for α, β and k with mean values corresponding to $\bar{\alpha} = 0.001\text{s}^{-1}$, $\bar{\beta} = 0.5\text{s}^{-1}$, and $\bar{k} = 100$ and the width $\sigma = 0.01$. In the top left corner of the plots the correlation coefficient is displayed. The bottom right plot shows the strong correlation between $\ln\alpha - \ln\beta + \ln k$ and h_{crit}. This Fig. is from [21]

while for $h > h_{\text{crit}}$, the steady state becomes unstable and a stable limit cycle shows up (s. Fig. 3.3). For the values $\alpha = 0.001\text{s}^{-1}$, $\beta = 0.5\text{s}^{-1}$, $k = 100$, we find $h_{\text{crit}} \approx 2.78$. Concentrations are measured in molecules per cell.

3.3 Distribution of a bifurcation parameter in a genetic network with uncertain parameters

The parameters α, β, and k were drawn from log-normal distributions such that $\log_{10}\alpha$, $\log_{10}\beta$, and $\log_{10} k$ are independent and normally distributed with a standard deviation of σ and mean values corresponding to $\bar{\alpha} = 0.001\text{s}^{-1}$, $\bar{\beta} = 0.5\text{s}^{-1}$, and $\bar{k} = 100$, respectively. We performed 10000 simulations for distributions with widths both $\sigma = 0.01$ and $\sigma = 0.2$. Drawing sets of parameters from these distributions we obtain instances of the dynamic system with undetermined Hill coefficient h. We then let the hill coefficient vary for each such realisation by running a bifurcation analysis to determine the critical value $h_{\text{crit}}(\alpha, \beta, k)$ of the Hill coefficient.

For a distribution width of $\sigma = 0.01$ we were able to determine the bifurcation point for each instance drawn from the distribution. Fig. 3.4 shows histograms of the random parameters α, β and k as well as the resulting histogram for h_{crit}. In 10^4 simulations,

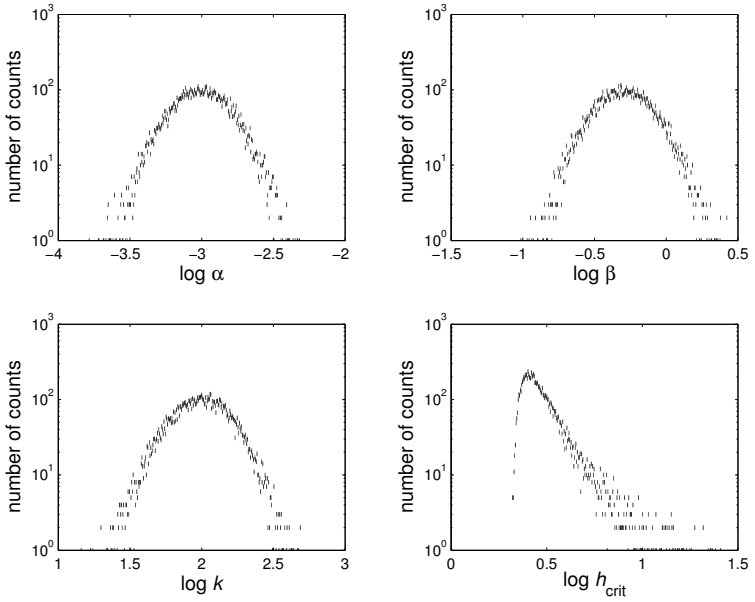

Figure 3.6: Three of the four plots show histograms of the distribution the parameters α, β and k were drawn from. The mean values correspond to $\bar{\alpha} = 0.001\text{s}^{-1}$, $\bar{\beta} = 0.5\text{s}^{-1}$, and $\bar{k} = 100$. The width was in each case $\sigma = 0.2$. The bottom right plot show the histogram for the resulting distribution of the value h_{crit}.

no critical value lower than ~ 2.68 was found. The correlation between the individual parameters and h_{crit} is shown in the (logarithmic) scatter plots in Fig. 3.5. We find positive correlation values for α and k, respectively, while β is negatively correlated with h_{crit}. This shows that a lower damping (small α) and a stronger coupling (high β or low k) between genes makes the system more prone to oscillations.

For a parameter width $\sigma = 0.2$ we found a bifurcation point in 9779 out of 10000 simulations. Figs. 3.6 and 3.7 show the the parameter histograms, the resulting histogram for h_{crit} and the correlations plots. The lowest value h_{crit} we found was ~ 2.09.

The qualitative behaviour of the cycle does not depend on the absolute scaling of time and concentration. This implies that h_{crit} can only depend on the linear combination $\ln\alpha - \ln\beta + \ln k$ which is confirmed by our simulation results (s. Figs. 3.5 and 3.7).

Parameter estimation for complex dynamic models is a challenge in current systems biology. To study the potential dynamic behaviour of a given model with uncertain parameters, we use a Monte-Carlo sampling approach. We draw the parameters from a distribution and observe the distribution of a property of the system.

In the example model describing a small genetic network, the occurrence of a Hopf bifurcation leading to qualitative change of the dynamic behaviour and the distribution of the connected quantity, the critical Hill coefficient, have been determined. We may also ask a slightly different question: if all parameters (including the Hill coefficients) are drawn from distributions, what is the probability for the system to oscillate? Given our distribution of h_{crit}, this can be easily answered by sampling h and h_{crit} independently

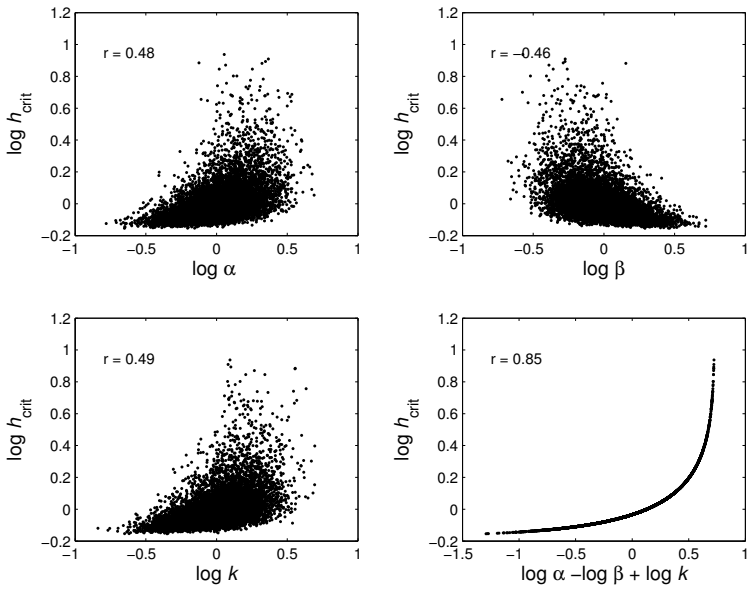

Figure 3.7: Correlation plots for the three parameters α, β and k versus h_{crit}. The log-normal distributions for α, β and k with mean values corresponding to $\bar{\alpha} = 0.001\text{s}^{-1}$, $\bar{\beta} = 0.5\text{s}^{-1}$, and $\bar{k} = 100$ and the width $\sigma = 0.2$. In the top left corner of the plots the correlation coefficient is displayed. The bottom right plot still shows that h_{crit} is a function of $\ln\alpha - \ln\beta + \ln k$.

from their distributions and counting how often $h > h_{\text{crit}}$.

The presented analysis can be considered as first step towards a thoroughly parameterized model. It gives hints about which types of qualitative behaviour can be expected at all and at what constellation of parameter values. It assesses which parameter values have a strong influence on the dynamics. This in turn is helpful in pointing to parts of the model where further more precise measurements are necessary. It clarifies where fluctuations might have a strong impact and they can be disregarded.

4 Collecting data for kinetic parameters

4.1 Parametrising models

The purpose of any model of a biological network is, first of all, to reproduce known facts about a biological system. As we saw in the last chapter, the values of the kinetic parameters of a model can have significant influence on the dynamical behaviour of a model. Thus, reproducing facts means finding the correct parameters such that the model behaves as expected. When during model construction kinetic parameters lack, time series of cellular components are used for parameter estimation, searching that set of parameters that best reproduces the time series.

But a satisfactory model will even go further and be used to make predictions about the system. These are then checked by experiments. Discrepancies between model and experiment, in turn, requires model revising followed by further experimental verification. This iterative cycle of model revision and experimental verification constitutes the cycle of mutual refinement of experimental and modelling knowledge about a biological system.

Model predictions concern properties of the systems that arise from the non-linear interaction dynamics of the model [11, 93]. Predictions include insight through simulation into the relative importance of components or modules in achieving a certain function. Furthermore, physiological malfunction or genetic defects can be mimicked by enhancing or weakening specific fluxes or rates associated with the respective proteins.

Model predictions and reproduction of known facts can only be achieved with correct parameter values. When possible, these values come from experiments that were undertaken under those conditions that the model is supposed to simulate. Another way of assigning values to kinetic parameters is searching the literature and electronical resources like BRENDA [134] for retrieval of kinetic data. These sources should also give the kinetic mechanism of interactions.

But in many cases information on the kinetic mechanism of an enzyme, in a specific organism or organelle, as well as the appropriate parameters are not available. Modellers then have to resort to other means and ways of parametrising their models, like making educated guesses.

As mentioned earlier, one way of dealing with the lack of kinetic parameters is using time series data of components of the system for parameter estimation purposes. Hereby, kinetic mechanisms have to be assumed unless further knowledge about the enzyme or protein complex can be taken into account. Another heuristic way of knowledge-based guessing of kinetic and cellular parameters is searching for information about the same enzyme or protein complex in closely related organisms [2].

Not only kinetic parameters, information about protein and metabolite concentrations

also is of use for metabolic modelling, because modelling with differential equations (s. Eq. (2.4)) requires initial values of the concentrations of the components.

4.2 Data sources of parameters

For the purpose of assigning numerical values to parameters of metabolic models we have thus collected three types of data from a variety of sources. First, data for cellular abundances or concentrations of proteins, mRNA and metabolites serve to set or reasonably guess initial values of concentrations in the model. Second, data of kinetic parameters are needed for the kinetics of the enzyme catalysed reactions. Third, thermodynamic data, impose constraints on the kinetic parameters and thus introduce dependencies between them (s. Sect. 2.5).

Concentrations of cellular components

(1) We take metabolite concentrations from Albe et al. [2]. In this publication concentrations of enzyme substrates are calculated from values for metabolite concentrations per weight of dry or fresh tissue and the water content of the respective fresh tissue, collected in a literature survey. The metabolite concentrations contain values for species covering three kingdoms: *Escherichia coli*, *Saccharomyces cerevisiae*, *Dictyostelium discoideum*, *Homo sapiens*, *Rattus sp.*, *Oryctolagus cuniculus*, *Vigna radiata*. The tissues are liver, muscle and heart for *rattus sp.*, red blood cells for *Homo sapiens* and *Oryctolagus cuniculus*, seeds for *Vigna radiata*.

For each concentration value the metabolite name and the organism as well as the tissue are given.

(2) We included three sources for protein abundance data. The first contains measured abundances. The two other publications, comparing protein abundance and mRNA expression for the yeast *Saccharomyces cerevisiae*, provide computed protein abundance data by integrating data for protein and mRNA abundance from several experimental sources.

Ghaemmaghami et al. [58] measured protein abundances by Western blot analysis for a TAP-tagged fusion library of the yeast *Saccharomyces cerevisiae* grown in rich medium to log-phase. The abundances are available at the Yeast GFP Fusion Localization Database [120] and are enhanced by localisation information [75].

Greenbaum et al. incorporate two-dimensional electrophoresis and multidimensional protein identification technology (so-called MudPit) data sets into their study [68]. They map each of the protein data sets independently into mRNA space by calculating a non-linear transformation onto a reference data set of mRNA abundances. By using the inverse of the calculated transformation of one of the protein datasets chosen by the authors, the resulting computed mRNA data sets are then transformed back into protein space (for details s. [67]).

Beyer et al. [16] combine the computed reference protein data set from [68] and the experimental data from [58] by taking the arithmetic mean if both data sets provide a value for an ORF, and the value from the respective source if only one is present. For the proteins that have a value in both data sets there often is considerable difference in order of magnitude between the values of the sources (s. supplemental material of [16]).

Each of theses sources gives the abundance value and the ordered locus name of the gene and, if existent, the gene name.

(3) Genome-wide mRNA abundance data for yeast wild-type cells, the yeast transcriptome, grown in YPD (Yeast extract peptone dextrose) medium under log-growth conditions without any stressors or chemical agents are found in [16, 73].

Holstege et al. measured the abundances with high-density oligonucleotide arrays (HDA) that allow for absolute measurements of abundances in the range from 0.1 to several hundred mRNA copies per cell [146].

Beyer et al. compile an mRNA expression data set from seven publications (s. supplemental material of [16]). Each data set is normalised to 15000 mRNA molecules per cell [146]. For each gene the reference value is chosen to be the median value among the data sets. Finally, the data are corrected for saturation (s. supplemental material of [16]).

In all sources the abundance value is associated with the ordered locus name (OLN) and, if existent, the gene name of a protein coding sequence.

Kinetic parameters

(4) K_M values are obtained from the public database BRENDA [134, 130, 131]. The K_M values in BRENDA are collected from the literature. At the top level all data in BRENDA are classified by the EC number but can also be searched by other criteria. Organism, substrate, if possible experimental conditions, isoenzymes and presence of activators, and a literature reference are given. In most cases a PubMed identifier [113] can be revealed behind the literature reference.

Enzyme naming complies to the IUBMB (International Union of Biochemistry and Molecular Biology) standards. Organism names are taken as they are from the literature sources if they lack systematic names. Reaction names are as defined by the IUBMB. Substrate names are those found in the literature.

(5) K_I values are obtained from the public database BRENDA [134, 130, 131]. They are complemented by the inhibitor name, organism name, a commentary about experimental and assay conditions. The literature reference is given.

(6) k_{cat} values are obtained from the public database BRENDA [134, 130, 131]. They come with information about the converted substrate, the organism, experimental conditions, as well as the literature reference.

Data for parameters and their sources

parameter	data source
concentrations of components	
metabolite concentration	[2]
protein abundance	[120, 16, 67]
mRNA abundance	[73]
kinetic parameters	
K_M values	[134]
K_I values	[134]
k_{cat} values	[134]
transcriptional frequency	[73]
mRNA half-life	[73]
thermodynamic constraints	
equilibrium constant	[63]
Gibbs free energy of formation	[3, 71]

Table 4.1: For modelling kinetic models of metabolic networks, different types of data serve to assign values to the parameters of a model. Three types of parameters come into play in kinetic modelling. First, concentrations of biomolecules serve to make reasonable assumptions about initial values of the biomolecules involved in the model, i.e. metabolites, mRNAs and proteins. Second, parameters for enzyme kinetics have to inserted into the kinetic expressions of the rates of each single enzyme catalysed reaction in the metabolic network. Third, thermodynamic data impose constraints onto the kinetic parameters throughout the whole metabolic network.

(7) Transcriptional frequencies for the yeast *Saccharomyces cerevisiae* come from [73]. They are calculated from steady state mRNA abundances and the half-lives determined from mRNA abundances in a temperature-sensitive mutant (s. (4.2).

(8) Half-life data for mRNA transcripts in the yeast *Saccharomyces cerevisiae* are available from [73]. They are computed from the mRNA expression levels in a temperature sensitive mutant strain of the yeast *Saccharomyces cerevisiae*, rpb1-1, 45min after a shift to a non-permissive temperature of 37°C. At this temperature the temperature-sensitive mutants immediately shut down the transcription of protein-coding genes and only degradation affects the mRNA levels. By comparing the mRNA levels in the mutant to those in wild type after 45mins, an apparent, first-order half-life is calculated.

Thermodynamic data

(9) Apparent equilibrium constants, $K_{eq}^{'}$, of biochemical reactions are available at the public database TECRDB(Thermodynamics of Enzyme-Catalysed Reactions) [63, 62]. The data in the database are retrieved from a literature search. Besides the reference itself, the following information is given with the apparent equilibrium constants if found in the reference source: reaction, enzyme, method of measure-

ment, conditions of measurement (temperature, pH, ionic strength, buffers, co-factors). The quality of the data are assessed and sometimes a commentary is added.

Apparent equilibrium constants K'_{eq} are used to calculate the transformed Gibbs free energies of reaction

$$\Delta G'^{0} = -RT\ln K'_{eq}$$

for biochemical reactions that account for sums over different ionised forms of a reactant [4]. When the pH and, in some cases, free concentrations of certain metal ions are given, it is the transformed Gibbs energy of reaction that is the criterion of equilibrium and that determines the apparent equilibrium constant [4]. The apparent equilibrium constant K'_{eq} for a biochemical reaction is written in terms of sums of species.

For the case of hydrolysis of ATP to ADP, the biochemical reaction occurring at a certain pH and pMg is expressed as

$$\mathrm{ATP} + \mathrm{H_20} \rightleftharpoons \mathrm{ADP} + P_i \tag{4.1}$$

with P_i denoting orthophosphate. In this reaction, ATP stands for a mixture of ATP^{4-}, HATP^{3-}, $\mathrm{H_2ATP}^{2-}$, MgATP^{2-}, MgHATP^{-} and $\mathrm{Mg_2ATP}$ at equilibrium. This holds for ADP analogously. The apparent equilibrium constant then is [4]

$$K'_{eq} = \frac{\mathrm{ADP\ P_i}}{\mathrm{ATP}\ c^0}\ . \tag{4.2}$$

with c^0 the standard state concentration of 1M.

The TECRDB follows the recommendations for biochemical thermodynamics by the IUPAC (International Union of Pure and Applied Chemistry) [4, 115] and the nomenclature of biochemicals by the IUBMB.

(10) Standard Gibbs free energies of formation of biochemical species are found in [3, 71]. They refer to biological standard conditions, namely a pH value of 7 and a temperature of 298.15K. Both references provide computed Gibbs free energies of formation.

4.3 Annotation of parameter data

The collected data for kinetic parameters and concentrations of cellular components are useful for modelling purposes only if each experimental value is fully annotated and can be associated with its meaning in the mathematical encoding of a model. A human modeller can then integrate information from different resources of knowledge to manually put together a model of a metabolic network.

But with systems biology the need for electronical exchange of biological knowledge, of experimental data for cellular components and their interactions, as well as of models

has grown. The various types of knowledge, the biological network information and experimental data, that enter a mathematical model of a biological network need to be linked to each other. There is a hierarchy in the way these heterogeneous types of information refer to each other.

At the basis of this hierarchy is the biological knowledge. In recent years electronical resources containing knowledge about the biological networks have been built. These pathway databases contain knowledge about cellular components and the networks they form. They represent inventories of the components of the cell as well as the structure of the cellular networks. KEGG (Kyoto Encylopedia of Genes and Genomes) is a prominent example of a pathway database [85, 99] linking genomes to pathways and including biochemical reactions. Other pathway databases are Reactome [140, 124] and BioCyc [87, 81]. ChEBI (Chemical Entities of Biological Interest) is a resource of biochemical compounds confining itself to the so-called small molecules in contrast to genome-encoded macromolecules and classifies them according to their biological function and their structure [42].

At the next level are resources containing the data for concentrations of cellular components and the kinetic data describing their interactions. Each data has to be associated with the knowledge that allows for its interpretation in terms of the biological network and its interactions. Electronic resources with data for biological networks are the BRENDA [134] or the Yeast GFP Fusion Localization Database [120].

Finally, dynamic models of biological networks refer to both resources of information, to the biological knowledge about components and interactions as well as to the data reflecting their dynamics. In order to achieve the electronical integration of data and interoperability of resources for modelling, shared standards for knowledge representation and nomenclature are needed and have to be agreed upon. The resources need to speak the same biological language.

To be of use in knowledge integration for the construction of metabolic networks, the data of concentrations of components and for enzyme kinetics have to be annotated with the appropriate information (s. Tbl. 4.2).

(1) Data of concentrations of cellular components have to identify the respective component and the organism they were measured in.

(2) As enzymes catalyse metabolic reactions, data for enzyme kinetic parameters necessarily have to refer to the enzyme, the reaction, and the organism they were measured in. Additionally, turnover rates k_{cat} are, in the case of a reversible process, meaningful only when the direction of the reaction they refer to is known. For K_m value data we also need to know the respective substrates. In the case of data for K_i values the inhibitor name is required.

(3) In the case of data for kinetic parameters describing mRNA transcription frequency and half-life as first-order processes, the ORF and the organism have to be denoted.

(4) Thermodynamic data merely regard biochemical reactions. Thus data for equilibrium constants refer to a reaction, and data for Gibbs free energies of formation concern metabolites.

Information required for data completeness

parameter	metabolite	enzyme	reaction	gene	organism
metabolite concentration	×				×
protein abundance				×	×
mRNA abundance				×	×
K_m values	×	×	×		×
K_I values	×	×	×		×
k_{cat} values [1]	×	×	×		×
transcriptional frequency				×	×
mRNA half-life				×	×
equilibrium constant			×		
Gibbs energy of formation	×				

Table 4.2: The data for model parametrisation are the dependent entities of the database. An entry in the respective table is complete only if additional information, as represented by the crosses (×), is associated with a sample of the respective data type. [1]The metabolite is used to identify the direction of the reaction.

4.4 Biological standards for model parameters

Data useful for parametrising models of metabolic networks are found in the literature and databases (s. Tbl. 4.1 in Sect. 4.1). These data, according to their type, are annotated with instances of one or several of the following biological concepts (s. Tbl. 4.2):

1. metabolite,
2. reaction,
3. enzyme,
4. gene,
5. organism.

In general, the annotation found in literature does not agree with the according biological information as stored in electronical resources. The main reason is that no consistent nomenclature is used.

For instance, names of biochemical compounds come in a variety of ways. Many substances have several common names that are traditionally used in biology. These can be ambiguous and are often used in various forms regarding, for instance, the inclusion of isomer and stereoisomer prefixes. The IUPAC has recommended a chemical nomenclature [114] that sometimes leads to unwieldy names and thus is not generally used in

practice. Resources like KEGG LIGAND or ChEBI, on the contrary, are dictionaries of small molecular entities in biochemistry giving not only a single name but also synonyms and spelling variants of the entities. The names are often the trivial names traditionally used in biology that also serve in describing enzyme reactions.

A knowledgeable human reader with the appropriate background will generally manage to link the information that comes with experimental data to electronical resources. But electronical assignment will in many cases not be sucessful. Mapping of the data from literature to the information stored in the electronical resources often encounters difficulties because string comparison requires exact matching. Stored synonyms, of course, enhance the chance of identification by string comparison.

Here we present an effort to map existing data useful for model parametrisation onto several reference resources chosen as standards. KEGG was chosen as a standard for naming compounds and reactions. Enzymes are classified according to the EC (Enzyme Commission) numbers [116]. ORFs are given with their systematic locus tags (also called systematic names, ordered locus names, OFR names) that are location-based systematic names assigned to genes in sequencing projects. Organisms are mapped onto the NCBI Taxonomy [55].

The data will thus be connected to and annotated with standard resources making them usable for computer tools and giving meaning to them within kinetic models of metabolism. Compounds and reactions from KEGG, EC numbers, locus tags from genome sequencing projects and organism names from the NCBI Taxonomy represent knowledge about the biological entities needed to describe processes of metabolism.

4.5 Standard references for model parameters

4.5.1 KEGG as a standard for metabolites and reactions

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a resource that was built to link genomes to biological function with the aim of reconstructing biological systems based on deciphered genomes [117]. The KEGG resource is an electronical representation of the molecular biology and biochemistry of the cell. Three of the databases making up the KEGG resource are: GENES, LIGAND, and PATHWAY.

(i) The database GENES contains complete and partial genomes generated from public resources, mainly the database RefSeq [125, 54] beside other sources like GenBank [15, 112] and EMBL [97, 80].

The collected genes are assigned KEGG orthology (KO) identifiers. The KO system is a categorisation of ortholgous genes that captures a gene's function based on pathways or protein networks [85, 86]. The KO consists of four levels. The top level contains five categories: metabolism, genetic information processing, environmental information processing, cellular processes and human diseases. The third corresponds to pathways and protein interactions, the fourth to genes. About one third to half of the genes contained in KEGG GENES have assigned a KO term [106].

(ii) The database LIGAND contains tables with key players of biochemistry: metabolites and enzymes and how these interact through biochemical reactions. They are stored in the databases COMPOUND, REACTION and ENZYME. Recently the database LIGAND was enhanced by tables for glycans and drugs. More specifically, KEGG COMPOUND is a chemical structure database for metabolic compounds and chemical substances relevant to biology. KEGG REACTION contains those reactions defined by the IUBMB (s. Sect. 4.5.2) as well as additional ones. For instance the enzyme alcohol dehydrogenase with EC number 1.1.1.1 is defined to catalyse the reaction `an alcohol + NAD+ <=> an aldehyde or ketone + NADH + H+`. But in KEGG, the enzyme 1.1.1.1 is associated with further reactions that actually appear in metabolic pathways with, for instance, the alcohol and the aldehyde specified as ethanol and acetaldehyde (R00754).

(iii) The database PATHWAY collects maps of networks of gene products or generalised protein networks. These generalised protein networks contain direct and indirect protein-protein-interactions [83, 84].

Direct protein-protein-interactions are found in protein complexes (e.g. RNA polymerase) or signalling pathways. Metabolic networks (mediated by the metabolites being shared between enzymes) and gene regulatory networks (mediated by the bound DNA producing its protein product) represent indirect protein-protein-interactions.

The five main categories of protein interaction networks in the PATHWAY database are metabolism, genetic information processing, environmental information processing, cellular processes and human diseases. The pathways are manually drawn based on literature and persistently updated according to biochemical evidence.

4.5.2 EC classification for enzymes

The EC (Enzyme Commission) nomenclature [142, 116] is chosen as a standard for enzymes. It classifies enzymes according to the catalysed reaction.

4.5.3 Ordered locus names for genes

As standard identifiers for genes ordered locus names (OLNs) are used.

In completely sequenced genomes generally not every gene is characterised and given a name. But during a sequencing project they are all systematically assigned OLNs that encipher the position of every, also the uncharacterised genes relative to each other in the chromosome space. They are not gene names but location holders identifying the locus.

OLNs are also called locus tags or systematic names. They are used in resources like RefSeq [54], SGD [123] or UniProt [36].

For the yeast *Saccharomyces cerevisiae*, for instance, an example of an ordered locus name is YLR113W. It denotes the gene with the name HOG1. The OLN encodes the location of an ORF in the following way [40]:

1. 'Y' stands for yeast.
2. The second letter, one out of 'A', 'B', ..., 'P', letter means one of the sixteen chromosomes.
3. The third letter, 'R' or 'L', stands for the right or left arm of the chromosome.
4. The three-digit number corresponds to the order of the ORF counting from the centromere.
5. The last letter, 'W' or 'C', designate the Watson or the Crick strand.

4.5.4 NCBI Taxonomy for organisms

The NCBI taxonomy database [55] is used as a standard reference for organisms. It contains more than 240 000 named organisms that are found with at least one sequence [144] in GenBank. Every new sequence submission to GenBank is checked for new organisms that are then classified and added to the taxonomy. The classification is based on phylogeny representing evolutionary relations as well as on morphological data.

4.6 Database entities

Once the data for parameters of kinetic models of metabolism are annotated with standard references, a database can be set up. Each parameter type (s. Tbl. 4.1) constitutes a table of the database. In addition, every biological concept used for parameter annotation stands for a table (s. Sect. 4.4). The tables containing the data for the parameters refer to the tables representing the biological concepts (s. Tbl. 4.2).

Databases are commonly represented by the entity-relationship (ER) model [27]. The information contained in a database is structured by the help of entities and relationships between them.

Entities represent abstract objects or concepts. They can be thought of as nouns. In a database they correspond to tables where each entry of the table is an instance of the respective entity. Usually the entries are said to be the rows of the table. The columns correspond to the attributes that define the concept.

For example, in the KEGG database the table REACTION contains instances of the concept 'reaction' one of which is the entry `R00200` with the definition `ATP + Pyruvate <=> ADP + Phosphoenolpyruvate`. Two attributes of this entry, the identifier and the biochemical definition, are mentioned here. Each entry (row) in the table REACTION has several attributes (columns) the values of which describe that specific instance.

Relationships, on the other hand, connect two or more entities and can be regarded as verbs. They correspond to a column in one table that contains references to another table. For the last example from KEGG REACTION, the entry identified by `R00200` has references to those entries in the table COMPOUND which have the names `ATP`, `Pyruvate`, `ADP` and `Phosphoenolpyruvate`, respectively. The verb that describes the

relationship between the tables REACTION and COMPOUND can be thought of as 'consumes or produces'.

Entities fall into two groups, independent and dependent ones. Independent entities do not rely on other entities to be identified and semantically complete. Instead, dependent entities can only be completely specified by reference to at least one other entity. For example, a reaction is identified only if the involved metabolites are so. Thus the entity reaction depends on the entity metabolite.

In an electronical resource containing data for model parameters, the biological concepts make up the entities the data refer to. Thus, the biological entities constitute the independent entities, gene being an exception. A gene, as a nucleotide sequence, is part of the genome of an organism and thus refers to the concept of organism.

The concepts corresponding to the parameters of kinetic models refer to the components and interactions of the biological network. These entities are thus dependent on the biological ones. For instance, a K_m value in the Michaelis-Menten formula (s. Eq.(2.18)), describing the turnover velocity of an irreversible biochemical reaction in the metabolism of an organism, is fully identified only if information on the substrate, the enzyme and the reaction as well as the organism is given (s. Tbl. 4.2).

4.6.1 Biological entities

The biological entities, covering the different levels from the genome to metabolism, onto which the data will be mapped are (s. Sect. 4.4):

1. metabolite,
2. reaction,
3. enzyme,
4. gene,
5. organism.

The entities metabolite and reaction are biochemical. The entities gene and organism belong to domain of evolutionary biology where organisms are genetically related to each other. The gap between these two realms is bridged by the entity enzyme which belongs to both owing to its being a gene product and a catalyst of a biochemical reaction.

The interconnections between the biological entities are drawn in Fig. 4.1. They can be captured by a phrase of the following type: The increased expression of gene YKL060C of *Saccharomyces cerevisiae* leads to higher levels of the enzyme Fructose-bisphosphate aldolase augmenting the flux through the catalysed reaction F1,6BP $\rightleftharpoons$ DHAP + GADP.

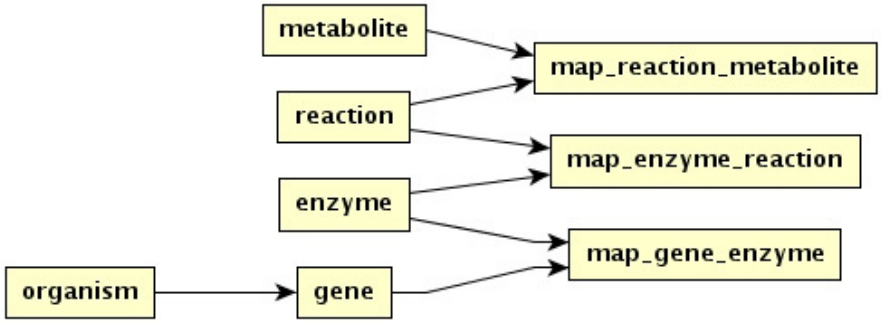

Figure 4.1: The biological entities as represented by their tables in the database and the interrelations between them. The biological entities cover genomic (entities organism and gene) and biochemical information (entities metabolite and reaction) that is linked by the entity enzyme.

4.6.2 Unit and reference entities

The data for parameters of kinetic models need not only to be annotated with biological knowledge. Each parameter type is measured in its unit. Furthermore, data for parameters refer to sources they were published in.

In order to simplify the structure of tables and their maintenance, we have scaled every data type to a single unit. In the literature data often come in a variety of units.

The data thus refer to two further entities:

1. unit,
2. reference.

4.6.3 Data entities

Each data type for parameters of kinetic models of metabolism amounts to a separate entity of the database (s. Tbl. 4.1). Each of these entities relates to one or several of the biological entities according to Tbl. 4.2.

1. metabolite concentration
2. protein abundance
3. mRNA abundance
4. K_m values
5. K_i values
6. k_{cat} values
7. transcriptional frequency

8. mRNA half-life
9. equilibrium constant
10. Gibbs free energy of formation

4.7 Database tables

The entities translate directly into tables of the database. Independent entities correspond to tables that do not refer to other tables. Tables standing for dependent entities contain columns with references to other tables.

4.7.1 Independent tables

1. `metabolite`,
2. `reaction`,
3. `enzyme`,
4. `organism`,
5. `unit`,
6. `reference`.

The independent tables correspond to concepts that are self-explanatory or explained elsewhere. Here the independent entities all refer to existing knowledge resources that were chosen as standards.

They are the following: KEGG serves as a standard for `metabolite` and `reaction`. The EC nomenclature is the standard for `enzyme`. OLNs (ordered locus names) used to denote ORFs in resources like RefSeq [54], SGD [123] or UniProt [36] denominate the entries of the table `gene`. The reference for `organism` is the NCBI taxonomy.

As to the non-biological entities, the table `unit` comprises standard SI units for each parameter type, and `reference` indicates the source of the data, either a literature or database reference.

4.7.2 Dependent tables

Among the dependent tables there is one biological one, namely gene. The other dependent tables contain the data to be used for model parametrisation. The entries in the corresponding table of each dependent entity will consist in a value for that quantity and references to the appropriate entries in the tables of the independent entities. The dependent entities are the following:

1. `gene`,
2. `metabolite_concentration`,
3. `protein_abundance`,
4. `mrna_abundance`,
5. `km`,
6. `k_mod`,
7. `kcat`,
8. `transcriptional_frequency`,
9. `mrna_half_life`,
10. `equilibrium_constant`,
11. `formation_enthalpy`,

where `k_mod` is designed to comprises both inhibition constants K_i and activation constants K_a.

4.7.3 Associative entities

The tables

1. `map_enzyme_reaction`,
2. `map_gene_enzyme`,
3. `map_reaction_metabolite`

constitute mappings between the entities `metabolite` and `reaction`, `gene` and `enzyme`, `reaction` and `metabolite`, respectively. These mappings are many-to-many meaning that a single entry in one table can refer to several entries in the other table, and vice versa. The information in these tables was extracted from KEGG.

In the case of reactions and enzymes, for example, there can be multiple forms of an enzyme that catalyse the same reaction like isoenzymes (derived from different genes). On the other hand there are enzymes that catalyse several reactions using either the same catalytic centre or, in the case of multienzymes, different parts or subunits.

4.8 Structure of Tables

Tables containing biological hierarchies

The tables of the basic (independent) biological entities have a simple structure (s. Tbl.(4.3)). They contain:

The biological tables of the database

table	reference id	reference name
`metabolite`	KEGG COMPOUND id	KEGG COMPOUND name
`reaction`	KEGG REACTION id	KEGG REACTION name
`enzyme`	EC number	official IUBMB name
`organism`	NCBI Taxonomy id	NCBI Taxonomy standard name
`gene` [1]	OLN	common name if existent

Table 4.3: The tables containing the biological knowledge used for data annotation all refer to chosen standards. These standards constitute inventories of the corresponding entity in the biological universe. Each instance fills a row of the table, the columns of the table contain the identifier or the name of the respective instance. For each entity the identifier and name of the biological object as given in the standard references were adopted here.

- an internal identifier as the primary key,
- an identifier from an external reference source taken as a standard; the KEGG identifier (in the case of `metabolite`, `reaction`), the identifier of the NCBI Taxonomy database [55] for `organism`, the EC number [18, 116] for `enzyme`,
- and the name given in these sources; i.e. the KEGG metabolite name is used for `metabolite`, the KEGG reaction string for `reaction`, the NCBI Taxonomy standard name for `organism` and the official name according to the IUBMB nomenclature for `enzyme`.

The data related independent entities have the following structure. The table `unit` has an internal identifier as primary key and columns for the unit name and the physical quantity it stands for. The table `reference` has, besides the internal identifier, simply a column for the reference name.

Tables containing data

The tables that contain the data for the kinetic parameters, concentrations and thermodynamic quantities all contain

- an internal identfier as the primary key,
- the value,
- possibly an error or standard deviation,
- references to the biological tables in order to interpret the data correctly and in a meaningful way (s. Fig.(4.2) and Tbl(4.1)).

Number of data entries inserted into tables

table	no. of entries	no. of original data
`metabolite_concentraton`	225	225
`protein_abundance`	10,141	10,141
`mrna_abundance`	12,267	12,267
`km`	47,449	76,427
`k_mod`	5,728	14,446
`kcat`	10,976	22,291
`transcriptional_frequency`	4,994	4,994
`mrna_half_life`	5,257	5,257
`equilibrium_constant`	2,088	3,409
`formation_enthalpy`	9,696	9,696

Table 4.4: Statistics of the data tables. The number of data entries For the enzyme kinetic parameters the number of values found by text mining is also given.

4.9 Constructing the database

The database containing the data for parametrising kinetic models of metabolic networks is set up with the MySQL database management system. Data were collected from the sources given in Sec. 4.1.

In order to insert the data into the database and make them electronically accessible, each single data point has to be annotated with the appropriate biological knowledge, represented as entries in the database. Every measured parameter has a meaning in the context of dynamical models of biological networks. What meaning it has in the interaction network is determined by its definition and is specified by naming the components of the biological network it relates to.

The biological information coming with the data has to be mapped to the biological information that is stored in the database chosen to be the naming standard. The mapping is done by comparison, by string comparison of the information in the source against that in the database.

If the data in the source are already annotated with standard names or identfiers there is no work to do anymore. For some of the data sources given in Tbl. 4.1 this is the case. Which information is required for annotation of which type of parameter shown in Tbl.(4.2).

1. Into the database table `metabolite_concentration` 225 metabolite concentrations were included. All concentration values were annotated with the appropriate entry in each of the tables `metabolite` and `organism`.

2. We entered more than 10,141 abundance values measured for 4,237 proteins of the yeast *Saccharomyces cerevisiae* into the table `protein_abundance`. As these data came with ordered locus names they were readily inserted into the database without any loss.

3. More than 12,267 mRNA abundances measured for 6,208 mRNAs of the yeast *Saccharomyces cerevisiae* could be entered into the database table `mrna_abundance`. As in the case of protein abundance data, the mRNA abundances had been annotated with standard ordered locus names and could be completely used for the database.

4. By far the largest amount of data for a parameter type is found for K_m values. These come from BRENDA [134]. 47,449 entries out of 76,427 in total were entered into the table `km` and could be mapped simultaneously onto an entry of the tables `enzyme`, `metabolite` and `organism`. Of these, 21,538 could also be assigned an entry of the table `reaction`.

   We also included 14,290 values from an effort of gaining knowledge about kinetic parameters from PubMed abstracts [70]. All of these are annotated by entries of the tables `enzyme`, `metabolite` and `organism`. Of these, 9,349 are also associated with an entry from the table `reaction`.

5. 5,728 K_i values from BRENDA [134] were inserted into the table `k_mod`. These could each be mapped to entries in the tables `enzyme`, `metabolite` and `organism`. 2776 of these could be mapped to an entry of `reaction`. The original data contain 14,446 K_i values.

   Text mining PubMed abstracts for kinetic parameters yields 5,069 entries of K_i values in `k_mod`. They are assigned entries of the tables `enzyme`, `metabolite` and `organism`.

   So far there are data only for K_i values and none for K_a values.

6. 10,976 k_{cat} values were inserted into the table `kcat`, each of them with annotations from the tables `enzyme`, `metabolite` and `organims`. Of these, 5,338 are also linked to a specified entry from the table `reaction`.

7. 4994 transcriptional frequencies for genes of the yeast *Saccharomyces cerevisiae* were entered into the database. These data were entirely inserted as the data source had been annotated with ordered locus names.

8. Into the table `mrna_half_life` 5,257 mRNA half-lives were inserted, one for each ORF of the the yeast *Saccharomyces cerevisiae*. The data had originally been annotated with ordered locus names and could thus be completely inserted into the database without any loss.

9. 2,088 values for equilibrium constants were filled into the table `equilibrium_constant`. The equilibrium constants cover 294 distinct reactions. The original data contain 3,409 equilibrium constants, so in 1,321 cases the original reaction string could not be mapped to one of reaction names from KEGG REACTION.

10. 9,696 standard free energies of formation for 9441 identifed metabolites are entered into the database. The original data could be entirely entered into the database as they were annotated with the appropriate entries of the table `metabolite`.

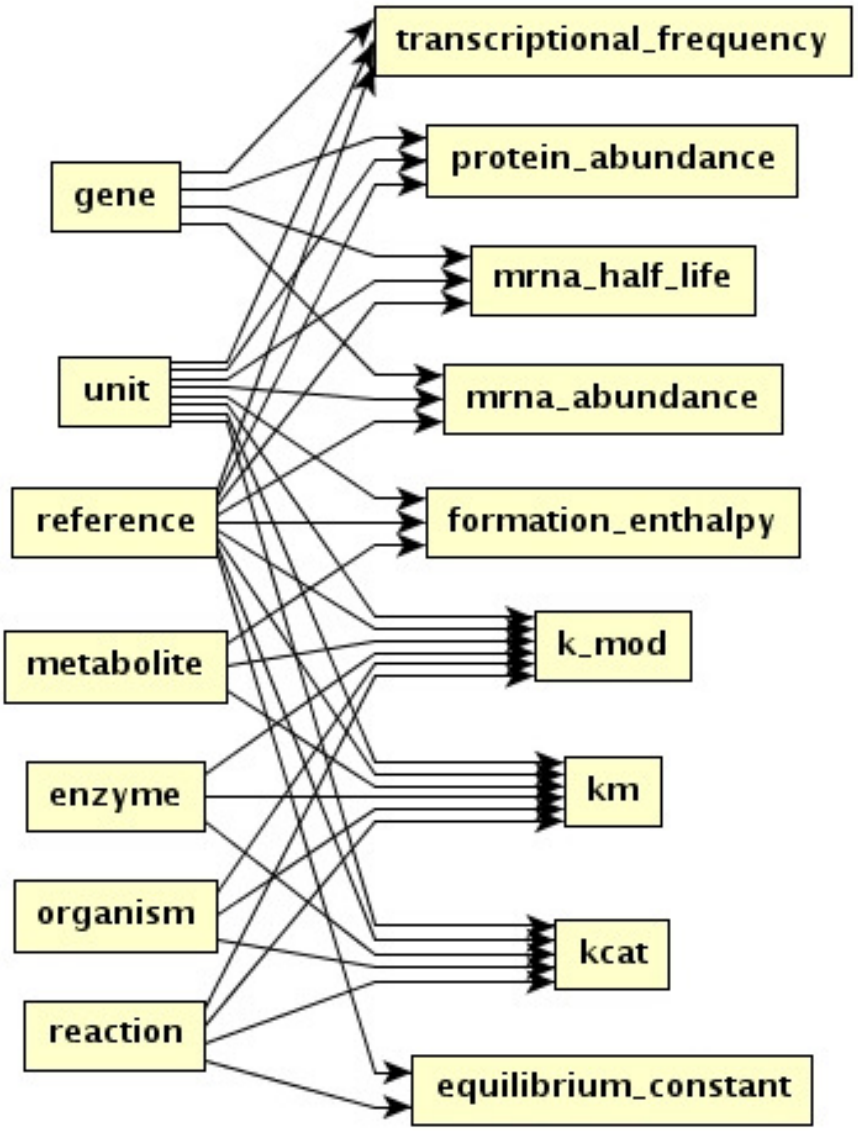

Figure 4.2: The structure of the database resulting from the relationships between the biological and the parameter entities. Entities and their tables are represented by boxes with the name inside. The entities that have arrows leaving the box are the independent entities: `metabolite`, `enzyme`, `reaction`, `organism`, `unit`, `reference`. The dependent entities are those with arrows pointing onto the respective box: `gene`, `map_enzyme_reaction`, `map_gene_enzyme`, `map_reaction_metabolite`, `metabolite_concentration`, `protein_abundance` , `mrna_abundance`, `km`, `kcat`, `k_mod`, `transcriptional_frequency`, `equilibrium_constant`, `formation_enthalpy`.

5 Automatically generated dynamical model of a metabolic network

In Chap. 4, we described the collection of data in an electronical resource useful for both model parametrisation and annotation of the data by standard knowledge references that are electronically accessible. The electronically stored data for parameters are amenable to computer tools.

We have set up a workflow [23] that automatically generates a kinetic model from the stoichiometric structure of a metabolic network. We begin either with the structure of a metabolic network, i.e. metabolites and reactions they participate in, or with a list of reactions alone from which we then build the stoichiometric structure. This core metabolic network, so far void of any dynamical information, then is fed into the pipeline. The pipeline assigns a kinetic formula to each reaction and then parametrises the entire network. The parametrised kinetics express the fluxes caused by the interactions and thus lead to a dynamical model.

The workflow, beginning with either a list of reactions or the stoichiometric structure of a metabolic network, consists of five main steps:

1. set-up of the structure of the metabolic network in the SBML file format [135],
2. assignment of a kinetic law to each reaction in the network,
3. collection of data for the parameters present in the kinetics of the reactions as well as of data for thermodynamic quantities that confine the kinetic parameters,
4. search of a set of parameters from the collected data accounting for the thermodynamic constraints,
5. output the parametrised model of the metabolic network in the SBML format.

The data for parameters that are stored in the resource generally stem from a variety of sources, were measured under different conditions and underly inherent as well as experimental fluctuations. The data at hand for model parametrisation thus are regarded as cues for model parameters, the true parameters being blurred.

Furthermore, thermodynamical constraints confine numerical values for model parameters. For this reason we choose a set of a parameters that is thermodynamically independent.

The set of parameter values that is finally inserted into the model is a function of the data for the parameters retrieved from the resource. From the extracted data we calculate the thermodynamically independent parameters. From these, in turn, we calculate the model parameters.

In order to distinguish between the various parameter types we will in the following refer to

- *data* on the one hand as those experimental or computationally derived values of parameters that stored in the electronical resource, denoted by x^*,
- *model parameters* as those values that are used for parametrisation of a kinetic model and
- *thermodynamically independent parameters* that can be varied freely without breaking thermodynamical constraints, denoted by θ.

5.1 Setting up the stoichiometric network

We base the structure of the network models on knowledge contained in KEGG [99]. KEGG is a set of databases that constitute a computer representation of biological knowledge at different levels, i.e. pathways, reactions, enzymes, compounds and genes. These levels are interconnected.

The metabolic networks are built from biochemical reactions stored in KEGG. We start with a set of reactions from the KEGG table REACTION. They are identified by the respective KEGG REACTION identifier or name. From the corresponding KEGG REACTION entries we retrieve information on the involved metabolites and enzymes. Thus, each interaction of the metabolic network is characterised by KEGG identifiers from the tables REACTION and COMPOUND and by an EC number. Those reactions that share metabolites are linked and constitute the connections of the network.

In the resulting SBML file encoding the stoichiometric structure of the network, each component of the network is identified by a MIRIAM-compliant annotations [100]. These contain Uniform Resource Identfiers (URIs) pointing to the respective external resource (s. MIRIAM Resources [43]) containing information on the respective network component. The data resources in the SBML file encoding the stoichiometry of the metabolic network are KEGG REACTION, KEGG COMPOUND and Enzyme Nomenclature.

5.2 Assigning kinetics to interactions

In order to turn the model that so far only encodes the stoichiometric topology into a dynamical model, each reaction is assigned a kinetic expression.

We use the convenience kinetics, a rate law that assumes a random-order enzyme mechanism and is applicable to reactions with any number of substrates, products and effectors [103] (s. Eq. (2.24)) for the meaning of the variables and parameters):

$$v = E_T \cdot \left(\prod_A \frac{\bar{c}_A}{\bar{c}_A + 1} \right) \cdot \left(\prod_I \frac{1}{\bar{c}_I + 1} \right) \cdot \frac{k^+_{cat} \prod_S \bar{c}_S^{\,n_S} - k^-_{cat} \prod_P \bar{c}_P^{\,n_P}}{\prod_s \left(\sum_{m=0}^{n_S} \bar{c}_S^{\,m} \right) + \prod_p \left(\sum_{m=0}^{n_P} \bar{c}_P^{\,m} \right) - 1} \quad . \tag{5.1}$$

Number of parameters per reaction required for convenience kinetics

Parameter type	number required
K_m	$N_S + N_P$
K_i	N_I
K_a	N_A
K_{cat}	2
E_T	1

Table 5.1: N_S: number of substrates, N_P: number of products, N_I: number of inhibitors, N_A: number of activators of a reaction.

This formula is directly applicable to any network model once its stoichiometric structure - encoded in the matrix N in Eq. (2.4) - and the regulatory structure - which effectors act upon an enzyme - are known. Convenience kinetics can thus be used by computer tools to turn a stoichiometric model into a dynamical one.

The number of kinetic parameters that is needed for modelling a network with convenience kinetics in Eq. (5.1) is: one K_m value for each substrate or product, one K_a value for each activator of the enzyme, one K_i value for each inhibitor and two turnover rates of the enzyme $k_{cat}^{\pm}$, one for either reaction direction, and the total enzyme concentration E_T. The total number of parameters needed per reaction is thus $N_S + N_P + N_A + N_I + 3$ as indicated in Tbl. 5.1 with N_S, N_P, N_I, and N_A denoting the number of substrates, products, activators and inhibitors of the reaction, respectively.

5.3 Retrieving the appropriate data for the parameters

The database containing data for parameters of kinetic models of metabolism which was presented in the Chap. 4 is used for model parametrisation. So far the model, encoded in SBML, is a mere stoichiometric structure of biochemical reaction containing no dynamical information about the enzyme mechanisms. The stoichiometric structure of the model is annotated with identifiers of standard references denoting the components of the metabolic network and the biochemical reactions.

This information about the components and biochemical interactions of the network is used to retrieve relevant parameters from the database. The metabolic network is likely to be modelled for a specific organism. For this organism the following information is searched in the resource.

1. KEGG COMPOUND identifiers are used to find concentrations of metabolites from the table `metabolite_concentration`.

2. EC numbers are used for collecting the respective protein abundances. As protein abundances are annotated with the appropriate ORF name of the gene that encodes it, further information is needed to link EC numbers and ORFs. This is achieved with information from KEGG ENZYME. Knowing which enzyme is encoded by

Properties of distributions of data in database

quantity	mean	std. dev.	5%-quant.	median	95%-quant.
metabolite concentration (mM)	1.419	4.496	0.003	0.123	5.15
protein abundance	10605	48174	279	2939	33502
mRNA abundance	3.33	13.85	0.1	0.76	10.0
K_m (mM)	15.50	392.10	0.001	0.14	20.005
K_i (mM)	11.9	173.0	$3 \cdot 10^{-6}$	0.016	13.7
k_{cat} (1/s)	1898	26757	0.008	6	1100
transcriptional freq. (1/hr)	7.4	19.6	0.3	2	28.9
mRNA half-life (min)	18.8	9.80	10	16	36
equilibrium constant	123.5	2303.4	10^{-6}	0.119	160.6
Gibbs energy of form. (kJ/mol)	-435.146	642.999	-1522.6	-331.0	324.3

Table 5.2: The parameter types collected in the database. Shown are properties of their distributions. All data show strong discrepancies between mean and median.

which gene, we can search for a protein abundance value for a given EC number for the table `protein_abundance`.

3. As kinetic parameters concern interactions more than one piece of information is required for their retrieval from the database. K_m values are looked for with a given EC number, the KEGG COMPOUND identifier of the substrate. If possible, the KEGG REACTION identifier will also be taken into account.

4. Thermodynamic data for the biochemical reactions of the metabolic networks are searched by KEGG COMPOUND identifiers in the case of Gibbs free energies of formation of metabolites and by KEGG REACTION identifiers in the case of apparent equilibrium constants.

5.4 Uncertain parameters

The collected experimental data x^* cannot be directly written into the kinetics of the reactions of the metabolic network. The found data x^* are uncertain and we calculate model parameters x from them by first finding a set of thermodynamically independent parameters θ. The method we adopt is represented by the scheme

$$x^* \longrightarrow \theta \longrightarrow x \,.$$

In the calculation of the model parameters x from the data x^* we will switch to the logarithms of the parameters and data. There are three reasons for this. Firstly, this makes the mapping between θ and x linear (s. below).

Secondly, as the quality and quantity of data is insufficient for consistent data for the model parameters we will construct probability density functions for the parameters. The probability distributions for the logarithmic parameters are approximated by Gaussian distributions, i.e. the actual parameters are log-normal variables. When we speak about logarithmic parameters we will stick to the names x, x^* and θ.

Properties of the distributions of the logarithmic parameters

quantity	mean	std. dev.	5% –quantile	median	95% –quantile
metabolite concentration	-0.92	1.07	-2.59	-0.91	0.71
protein abundance	3.45	0.62	2.45	3.47	4.53
mRNA abundance	-0.08	0.61	-1	-0.12	1
K_m predicted	-0.43	0.61	-1.35	-0.42	0.47
K_m from literature	-0.86	1.34	-3.01	-0.85	1.3
K_i	-1.95	2.06	-5.51	-1.8	1.14
k_{cat}	0.66	1.59	-2.1	0.78	3.04
transcriptional frequency	0.36	0.6	-0.52	0.3	1.46
mRNA half-life	1.24	0.17	1	1.2	1.56
equilibrium constant	-1.4	2.42	-5.99	-0.92	2.21
Gibbs energy of formation	2.14	0.54	1.13	2.21	2.89

Table 5.3: Shown are the properties of the distributions of the logarithmic parameters from the database. The means and medians are all close or very close except for the distribution of the equilibrium constants. For metabolite concentration, protein abundance and the K_m values the 5%-quantile and the 95%-quantile are symmetrically located around the median.

Thirdly, several data entries for a single parameters can be found in the database. And these can be considerably different in their order of magnitude. Averaging of logarithmic values of data copes better with large differences in the order of magnitude.

Uncertainties in the data x^* are manifest in different ways.

(i) Experimental data are noisy because of measurement errors. Sources of noise are biological variability, measurement errors and in vitro measurements in non-appropriate conditions.

(ii) Many parameters will not be available and therefore remain undetermined.

We nevertheless try to assign values to K_m values when data do not exist. When the database does not contain a K_m value present in a network model we make a guess based on knowledge about K_m values for the same enzyme in different organisms and for different enzymes in the same organism. We guess the missing K_m value based on a statistical linear model [22].

The idea behind is that there are different factors that explain a K_m value. The first of the three factors is the substrate contribution μ determined by the substrate's chemical properties that determine its binding to enzymes. Second, we presume conservation across organism of properties in the enzyme-substrate binding leading to factor α determined by the enzyme. The third factor β is the contribution from the organism assuming that enzyme adapts to typical substrate concentrations in the cell as it is the ration of the substrate concentration to the K_m value that determines the saturation of the enzyme. The logarithm of the K_m value is the $\ln(K_m) = \mu + \alpha + \beta$. This calculated guess of a missing K_m value is then joined to the set of collected data x^*.

(iii) We may find different different measured values, possibly also different orders of magnitude for the same parameter. In case more than one data for a specific parameter are found all data are taken into account and their contribution in the calculation of the set of model parameters x is weighted.

(iv) Thermodynamics imposes constraints on possible kinetic parameter values [103] as strict adherence to Haldane relationships must take place (s. Sect. 2.5), be it by measurement or estimation. The thermodynamical constraints only allow for certain combinations of parameter values. In fact, we choose a set of independent parameters θ and calculate the complete set of parameters x from it.

For a network modelled with convenience kinetics the following Haldane relationship confines the parameter values [103]:

$$K_{eq} = \frac{\prod\limits_{P} (c_P)^{n_P}}{\prod\limits_{S} (c_S)^{n_S}} = \frac{k_{cat}^{+} \prod\limits_{P} (K_{mP})^{n_P}}{k_{cat}^{-} \prod\limits_{s} (K_{mS})^{n_S}} \; . \tag{5.2}$$

In order to ensure thermodynamic consistency of the parameter set, we regard the turnover rates $k_{cat}^{\pm}$ as dependent on each other. We choose the geometric mean of them to be independent and call it the velocity constant $k_V = \sqrt{k_{cat}^{+} k_{cat}^{-}}$. With independent parameters K_{eq}, K_{mS}, K_{mP} and k_V, both k_{cat}^{+} and k_{cat}^{-} can be calculated from Eq. (5.2) (for details s. [103]).

Going further and we express the equilibrium constant K_{eq} through Gibbs energies of formation of the metabolites involved in the reaction according to Eqs. (2.29) and (2.30). This is useful when data of Gibbs free energies of formation for the reacting metabolites are available. With Eq. (2.30) the Haldane relationship Eq. (5.2) then becomes

$$e^{-\Delta G^0 / RT} = \frac{k_{cat}^{+} \prod\limits_{P} (K_{mP})^{n_P}}{k_{cat}^{-} \prod\limits_{S} (K_{mS})^{n_S}}$$

with (s. Eq. (2.29))

$$\Delta G^0 = \sum_{P} n_P \mu_P^0 - \sum_{S} n_S \mu_S^0 \; .$$

If we take the logarithm of the Haldane relationship (5.2) we get a linear dependency of the logarithmic parameters among each other. The same holds for the general Haldane Eq. (2.34) and for Eq. (2.29). We thus establish a linear relationship

$$x(\theta) = R_\theta^x \, \theta \tag{5.3}$$

between the independent parameters θ and the dependent parameters x [103], where the matrix R_θ^x is completely determinded by the network topology.

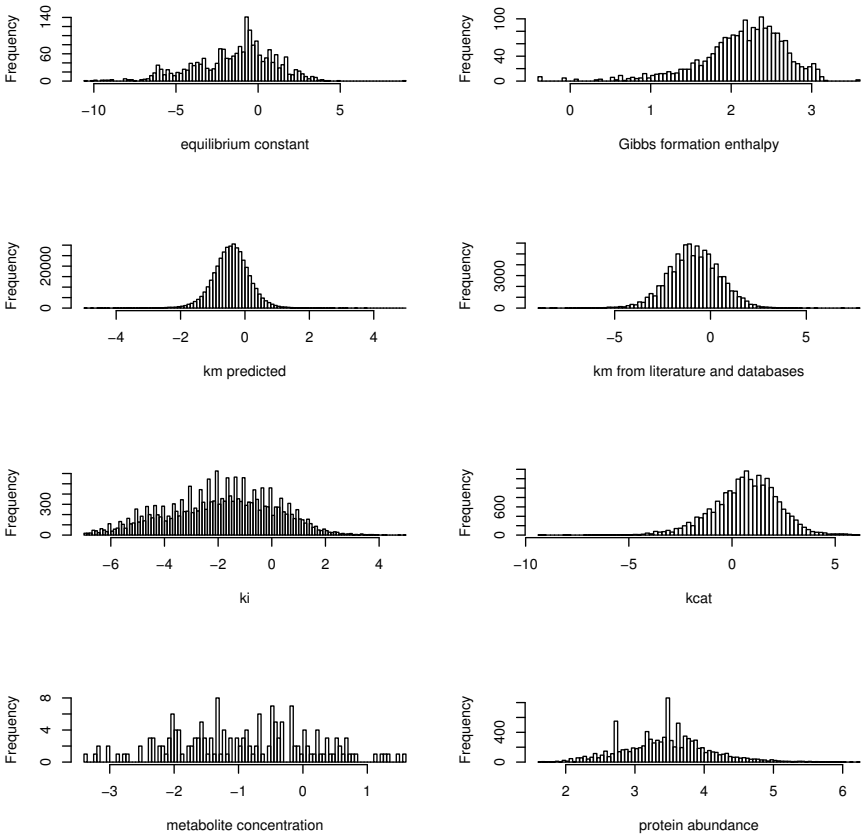

Figure 5.1: Prior distributions of logarithms of different data types used in the Bayesian parameter estimation approach.

We now relate the logarithms of the parameter data x^* retrieved from the electronical resource to the logarithms of the true parameters values x by

$$x^* = x + \epsilon \tag{5.4}$$

where ϵ is the uncertainty concerning the data.

Given collected parameter data x^* we estimate the independent parameters $\hat{\theta}$ from Eq. (5.3) by error minimisation. This corresponds to a maximum likelihood estimation with the likelihood function

$$L(\theta) = p(x = x^*|\theta)$$

and the estimated independent parameters

$$\hat{\theta} = \max_{\theta} L(\theta) \ .$$

We then plug the estimates back into Eq. (5.3) to yield the full set of (thermodynamically dependent) parameters x.

In order to account for already existent knowledge about the independent parameters θ we further employ Bayesian inference to determine the posterior probability distribution of θ given x^*:

$$p(\theta|x = x^*) \propto p(\theta)\, L(\theta)\,. \tag{5.5}$$

The prior distributions $p(\theta)$ for the independent parameters θ representing general knowledge about the parameters we obtain from the information in the database. These prior distributions $p(\theta)$ we derive from a statistics over all collected data for each parameter type. The properties of the distributions of the data in the database are indicated in Tbl. 5.2, for the logarithmic data in Tbl. 5.3. The distributions of some of the logarithmic parameters are also shown in Fig. 5.1.

For instance, a log-normal distribution fitted to the distribution of K_m values in the database Brenda [134] is used as a prior for each K_m value in the network (in Fig. 5.1 the K_m designated as from literature and databases).

With the posterior distribution from Eq. (5.5) different parameter set x can be calculated from different sets of thermodynamically independent parameters θ. The x are then inserted into the model and its dynamical properties can be assessed by simulation. By random sampling from the posterior distribution $p(\theta|x = x^*)$ we can find distributions of the behaviour of the model [21, 102].

5.5 Kinetic model

Once the entities of the metabolic network, i.e. reactions, metabolites and enzymes, are assigned their parameters, the result is written to an output file in SBML format. The annotations in the SMBL file accord with the MIRIAM standard [100, 43]. The dynamical model can then be simulated with any tool that is able to read SBML files.

5.6 Application to Sulfur-Methionine-Pathway in *Saccharomyces cerevisiae*

As a test case, we applied the described automatic model generation to the sulphur assimilation and the glutathione biosynthesis pathways in the yeast Saccharomyces cerevisiae. These pathways play an important role in the buffering of arsenic in order to avoid toxic effects: the cell increases the uptake of sulphur, leading to a raised glutathione level. Glutathione, having a high reduction potential, forms a complex with arsenic and the complex then is disposed in the vacuoles. The expression of the enzymes involved in these pathways is enhanced upon exposure to arsenic [137].

From a manually sketched metabolic network of the sulphur assimilation and the glutathione biosynthesis pathways, an enhanced version of the model in [137], we looked up the KEGG reaction identifiers. With these identifiers, information about the reactants

Statistics of data retrieved for model.

quantity	*S. cerev.*	number of data found for Glutathione synthesis
protein abundance	10141	13 of 15 enzymes
metabolite conc.	30	7 of 36 metabolites
K_m	2475	30 of 73 enzyme-metabolite pairs
$k_{cat}^{\pm}$	144	0 of 16 reversible reactions
equilibrium constant	-	2 of 16 reactions
Gibbs energy of form.	-	20 of 36 metabolites

Table 5.4: The number of data of different parameter types in the database, how many of them apply to Saccharomyces cerevisiae and how many have been extracted for the model (not only kinetic parameters).

and enzymes is fetched from the KEGG database. The result is the metabolic network shown in Fig. 5.2. We have so far not included any inhibitions or activations.

In the data retrieval step we could find 130 entries some of them refer to the same parameter of the model and are averaged. Among the data are 34 protein abundances, 7 metabolite concentrations, 30 K_m values, 26 K_i values, 2 equilibrium constants and 53 Gibbs free energies of formation. After averaging data belonging to the same component or interaction of the network we are left with 48 parameters: 13 protein abundances, 7 metabolite concentrations, 29 K_m values, 2 equilibrium constants and 20 Gibbs free energies of formation. No k_{cat} value could be found for this network (s. Tbl. 5.4).

Thus, most of the K_m values are missing and a large part of the metabolites no Gibbs free energies of formation are available. This are not even enough data to completely determine the model thermodynamically. Or stated otherwise, with the fetched data x^* not even all the thermodynamically independent parameters θ are assigned numerical values.

But all therodynamically independent parameters θ need to be ascribed values in order to calculate a complete set of model parameters x. As long as no other data are available we will set the parameter type to the mean value of the respective prior. This way we get a completely parametrised model.

The prior distributions of the logarithms of the parameters are shown in Fig. 5.1. In some cases they can be well described by a normal distributions (i.e. the parameter itself is log-normal distributed). Especially where the number of collected data (s. Tbl. 5.2) is large the normal distribution is a good description of the respective actual distribution. This is the case for the K_m, K_i, k_{cat} values and the Gibbs free energies of formation.

The kinetic parameters of the model are determined by 144 kinetic and thermodynamic values. Those parameters for which no data can be extracted from the database are mainly determined by the mean values of their prior distributions (s. Tbl. 5.3) and then undergo the thermodynamical adjustment and Bayesian procedure.

In Figs. 5.3 and 5.4 we show, for an example, the K_m values of the model. In the top figure we display the K_m values extracted from the database (missing data are indicated by grey diamonds with black borders) and their numerical values. In the bottom figure we display the model parameters after the replacement of missing values and the adjustment

to thermodynamical constraints in the course of the Bayesian procedure. High numerical values tend to stay high, missing ones are replaced by "average" numerical values.

When simulated with initial concentration values of the metabolites in the range of 0.1mM to 10mM, and holding the concentrations of the coenzymes constant, the model yields concentrations in the range of 1μM to 1mM. The fluxes obtain values in the range of 1nM/s to 1μM/s.

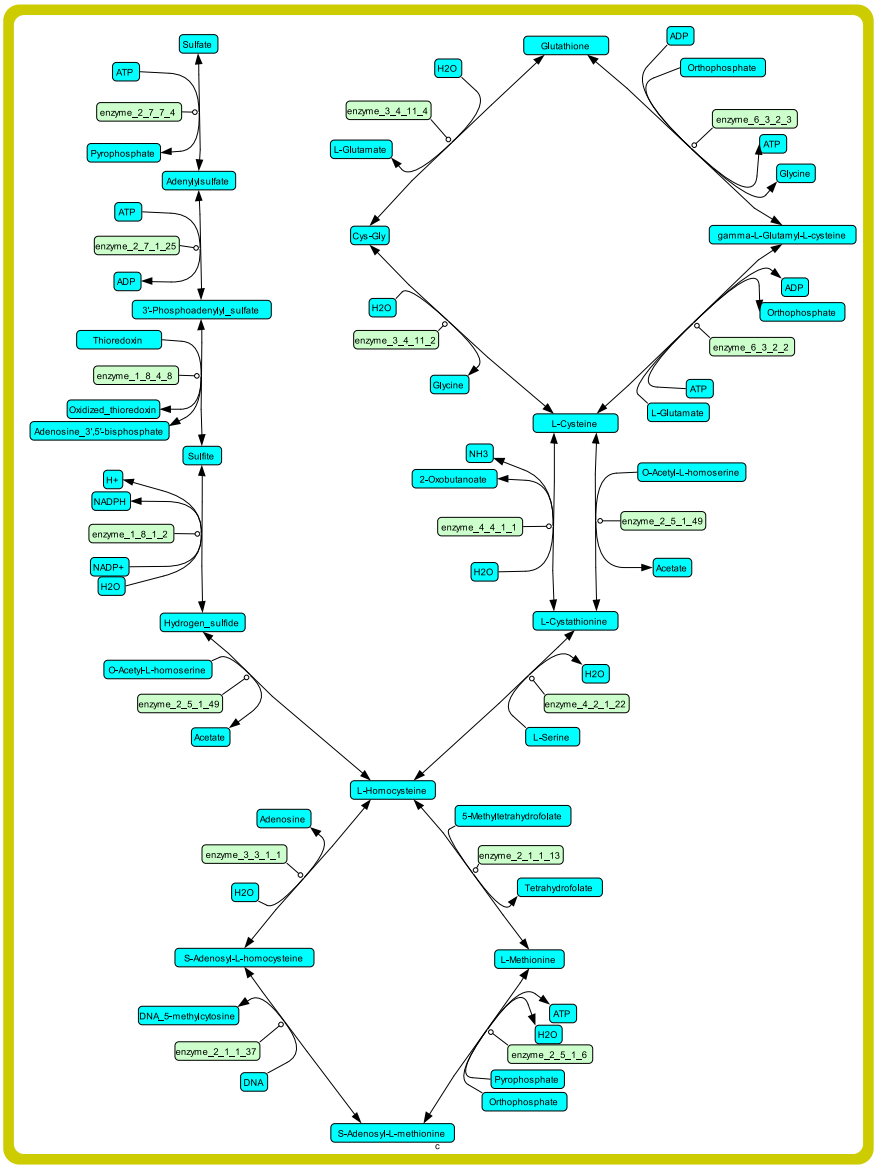

Figure 5.2: An automatically generated metabolic network set up starting with the KEGG REACTION identifiers R00529, R00509, R02021, R00858, R01287, R00192, R00380, R00177, R00946, R01290, R01001, R03217, R00894, R00497, R00494, R00899.

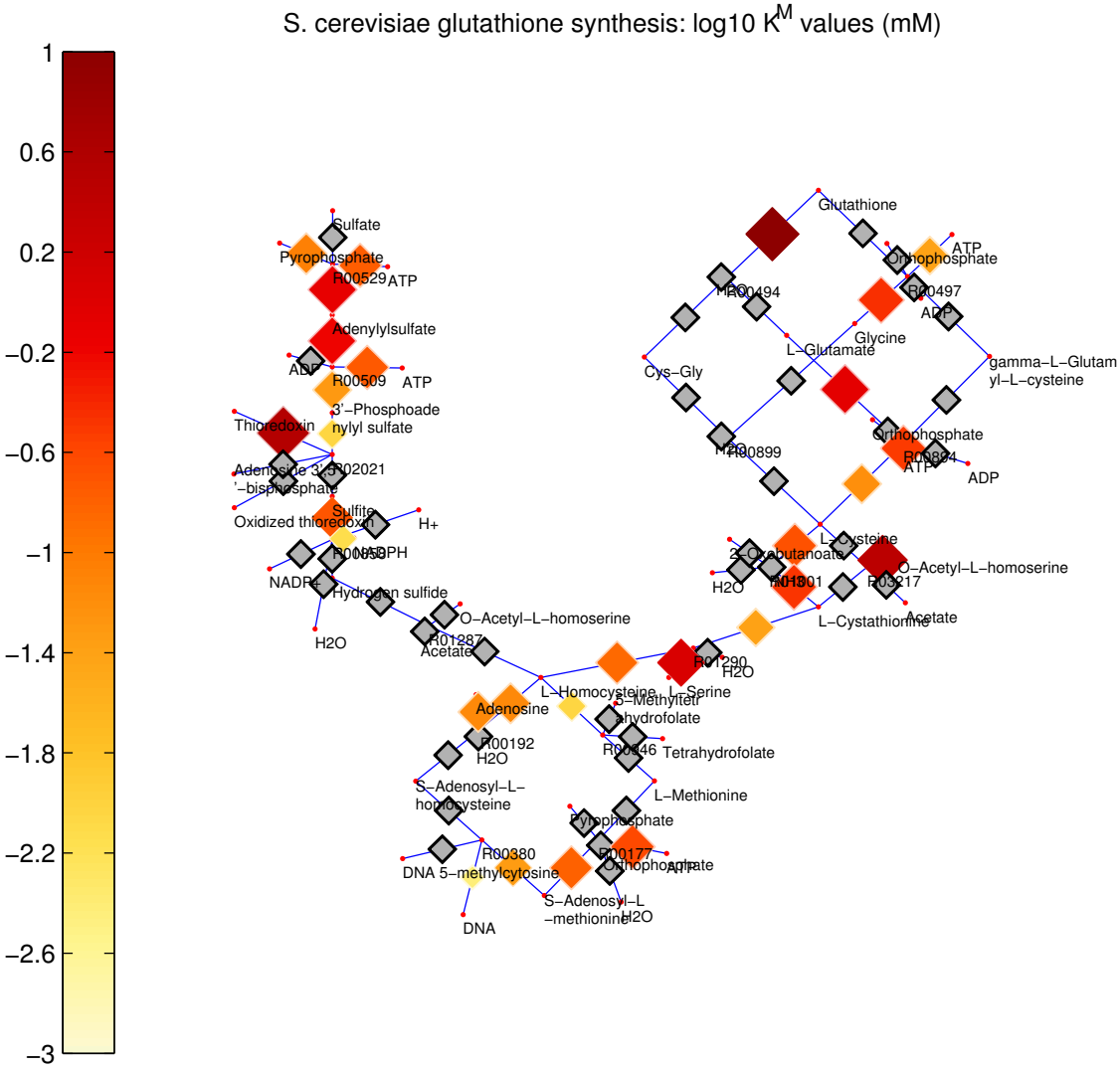

Figure 5.3: Michaelis-Menten constants in the sulphur-glutathione model. In this figure the K_m values for the model of sulphur metabolism retrieved from the database are displayed. The colour bar to the left indicates the logarithm of the respective K_m value measured in mM. The K_m values that are not assigned a value after the database search are shown in grey.

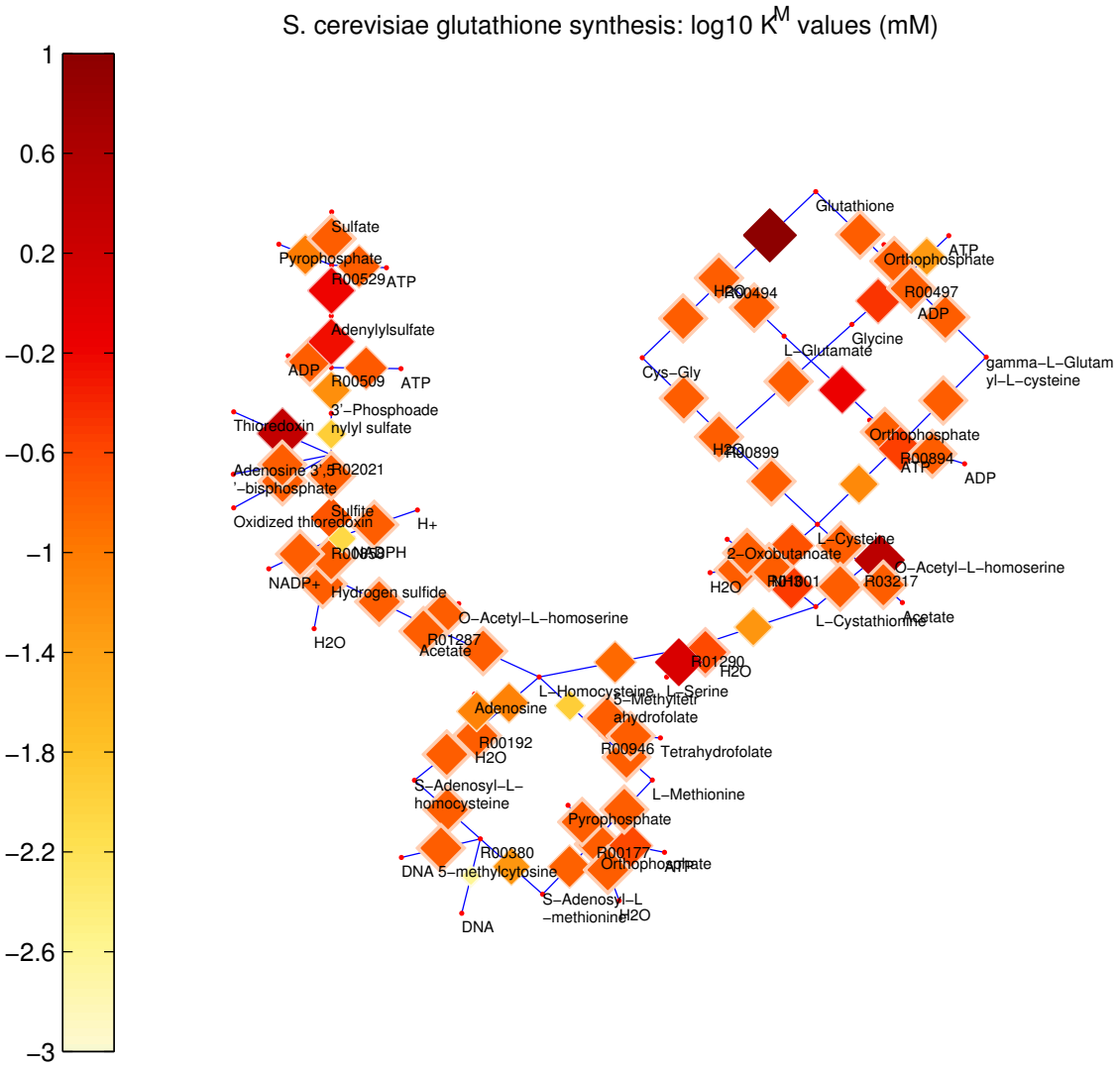

Figure 5.4: In this figure the K_m values are displayed after the model has been fully parametrised in a thermodynamically consistent manner. Those K_m parameters for which no data are available, indicated by grey diamonds in Fig. 5.3, are set to an average value. K_m parameters for which data were retrieved from the database, indicated by colored diamonds in Fig. 5.3, tend to keep their values. The colour bar to the left indicates the logarithm of the respective K_m value measured in mM.

6 Discussion

Systems biology seeks to gain an understanding of biological networks by modelling them dynamically. Ultimately this means revealing the proteins and their interaction networks in the cell.

Interactions of proteins are at the basis of any process in biology. They entail changes in protein's abundances in space and time. They determine a protein's function.

The identification of protein interactions in the cell is a formidable task. A wealth of techniques for discovering interaction of proteins with proteins or with other molecules like nucleic acids exists. A recent area of research, called chemical genomics, screens small molecules for the binding and modulating properties with respect to proteins.

But system biology requires knowledge not only about which network components interact. Systems biology needs dynamical knowledge about those interactions. In order to model interaction networks parameters need to be known that determine at which rates the interactions occur.

In Chap. 2 we give a brief introduction into the modelling of metabolic networks. Deterministic modelling of the dynamics of metabolism is based on rate equations. These require parameters that have to be measured. Time scale separation can save computational cost by concentrating on the processes that act on a particular time scale. This was shown with the example of Michaelis-Menten kinetics. Furthermore convenience kinetics were introduced that represent a random-order enzyme kinetics suited for automatically assigning kinetics to a reaction. We also point to the importance of thermodynamic consistency of kinetic parameters in a metabolic network.

If the vision of an *in silico* cell is to come true the interactions of, in principle, every single protein need to be quantitatively characterised. But the fact is, that to date not even models of subsystems can be fully and consistently parametrised. Data do either not exist or quality of data is poor. Indeed, discrepancies between measured fluxes and metabolite concentrations, on the one hand, and calculated ones from a model with measured kinetic parameters have been reported [136]. Parameters are either not known or of doubtful quality because *in vitro* enzyme assays often do not reflect *in vivo* conditions. Modellers mostly have to resort to parameter estimation. Altogether, knowledge about kinetic parameters is uncertain.

In Chap. 3 we examined the question of how uncertainties in the parameters of a kinetic model can entail different dynamical properties of a system. Specifically, we chose a system called the repressilator [45] which is a genetic network of three genes inhibiting one another such that a cycle forms.

This network either tends to a steady state or exhibits persistent ocscillations depending on the shape of the signal-response curve describing the interaction between the genes. With steeper signal-response sustained oscillations get more likely. This depends

on a single parameter, namely the Hill coefficient. This kind of assessment of possible dynamical properties of a dynamical network is helpful in determining where in the network further experimental investigation could be critical to increase the knowledge about a system.

With the advent of systems biology the need for electronically stored information is growing in order to exchange the rapidly increasing amounts of data and information. Electronical resources for different kinds of knowledge have been constructed. Those like KEGG or RefSeq represent biological knowledge. Others contain experimental data like ArrayExpress or GEO. BRENDA is a resource for information on enzyme kinetics.

For systems biology it will be important to link these different knowledgebases, to join for instance kinetic knowledge in BRENDA with biological knowledge in KEGG. So far, BRENDA categorises its knowledge according to the EC nomenclature and uses systematic organism names. But, for instance, substrate names are usually taken from the literature sources.

In Chap. 4 we present an effort to collect data about biological networks and to store them in a repository that makes them not only human readable but accessible to machines and automated handling. We have collected various data types.

Data for abundances of networks components like proteins or mRNAs are relatively easy to manage: each data value is annotated with the appropriate network component and possibly further explanatory information like the organism and experimental conditions. But data regarding interactions of network components are more involved to annotate because it takes at least the two interacting components to identify.

We undertook the effort of connecting kinetic data stored in BRENDA with data from KEGG. We compared information used to annotate data in BRENDA with the respective knowledge stored in KEGG by string comparison. The main difficulty was to coalesce substrate names in BRENDA with metabolites in KEGG. The reason is the lack of any adopted standard nomenclature for small molecules. The difference of a prefix in the names can already render impossible to join what a human readily joins. A lot of BRENDA data were lost because the substrate names could not be mapped.

In Chap. 5 we introduce a pipeline to automatically set up dynamical models of metabolic networks. This pipeline uses the database containing abundance, interaction and thermodynamic data presented in Chap. 4. It constructs a stoichiometric network and searches data for kinetic parameters, component abundances and thermodynamic quantities after every reaction has been assigned a kinetics according to the convenience kinetics scheme.

Convenience kinetics are readily assigned in an automated fashion to any biochemical reaction including effectors. It represent a random-order mechanism and remedies the lack of knowledge about enzyme mechanisms or the electronical inaccessibility of this knowledge.

The data assigned to the kinetics are scarce, mostly of poor quality and seldom thermodynamically consistent. We thus take the data, fill the gaps of missing data and infer a fully parametrised, thermodynamically consistent kinetic model in a Bayesian approach.

In order to achieve a fully parametrised kinetic model a minimum set of independent

parameters needs to be assigned values. But the data stored in the database are not numerous enough to do the job. In order compensate for these gaps and parametrise the whole network we have a model for predicting missing K_m values based on present data. For other data types the information in the prior distribution has to suffice as long as not more data are available or, for instance, BRENDA data are not better linked to resources like KEGG. Setting up a standard nomenclature for these so-called small molecules is the necessary step to resolve this problem. ChEBI, for instance, is a dictionary tackling this problem.

The constructed kinetic models are models in progress. Whenever more knowledge is available it can be included in the parametrisation. The models will thus hopefully become more realistic in the course iterative improvements. For instance, time course data of metabolites could be incorporated. They would represent the data and the distributions from the first step the prior information. One day we will hopefully have access to sufficient data of quality to model biological networks realistically.

Derivation of Michaelis-Menten kinetics by time scale separation

A classical example for time scale separation is the minimal metabolic network

$$\left(S + E \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} ES \right) \overset{k_2}{\rightarrow} P \tag{.1}$$

consisting of one irreversible enzyme catalysed reaction turning over species S into species P where E denotes the enzyme and ES the complex of enzyme and substrate.

The simplification in the dynamic description after time scale separation leads to Michaelis-Menten-kinetics. The fast process in this network is the binding of the substrate S to the enzyme E

$$S + E \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} ES \,, \tag{.2}$$

the process of interest is the production rate of product P

$$ES \overset{k_2}{\rightarrow} P \,, \tag{.3}$$

the slow process is the change of total enzyme concentration $E_T = E + ES$ because gene expression and protein degradation act on longer time scales. This leads to

$$\frac{dES}{dt} = k_{+1} \cdot S \cdot E - (k_{-1} + k_2) \cdot ES \tag{.4}$$

$$\frac{dP}{dt} = k_2 \cdot ES \tag{.5}$$

$$\frac{dE_T}{dt} = 0 \,. \tag{.6}$$

We solve the fast variable condition from Eq.(.4)

$$0 = k_{+1} \cdot S \cdot E - (k_{-1} + k_2) \cdot ES$$

for ES in order to insert it into Eq.(.5), thereby using the fact that the slow variable

remains constant (Eq.(.6)) by substitution of $E = E_T - ES$. We find

$$ES = E_T \frac{S}{\frac{k_{-1}+k_2}{k_{+1}} + S} \tag{.7}$$

$$= E_T \frac{S}{K_m + S} \tag{.8}$$

where the Michaelis-Menten constant is defined as $K_m = (k_{-1} + k_2)/k_{+1}$. This leads to following rate equation for the product P:

$$\frac{dP}{dt} = k_2 E_T \frac{S}{K_m + S} \tag{.9}$$

and the reaction scheme (.1) can be replaced by the following scheme

$$S \overset{(E_T, K_m, k_2)}{\longrightarrow} P \,. \tag{.10}$$

Bibliography

[1] E.K. Ainscow and M.D. Brand. Errors associated with metabolic control analysis. Application Of Monte-Carlo simulation of experimental data. *J Theor Biol*, 194 (2):223–33, 1998.

[2] K.R. Albe, M.H. Butler, and B.E. Wright. Cellular concentrations of enzymes and their substrates. *J Theor Biol*, 143(2):163–95, 1990.

[3] R.A. Alberty. Equilibrium Compositions of Solutions of Biochemical Species and Heats of Biochemical Reactions. *Proceedings of the National Academy of Sciences*, 88(8):3268–3271, 1991.

[4] RA Alberty. IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN). Recommendations for nomenclature and tables in biochemical thermodynamics. Recommendations 1994. *Eur J Biochem*, 240(1):1–14, 1996.

[5] R. Apweiler, A. Bairoch, et al. The human proteomics initiative (HPI). *Trends Biotechnol*, 19(5):178–181, 2001.

[6] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, 32:D115, 2004.

[7] R. Apweiler, A. Cornish-Bowden, J.H.S. Hofmeyr, C. Kettner, T.S. Leyh, D. Schomburg, and K. Tipton. The importance of uniformity in reporting protein-function data. *Trends Biochem Sci*, 30:11–12, 2005.

[8] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[9] W.T. Astbury. Molecular Biology or Ultrastructural Biology? *Nature*, 190(4781): 1124, 1961.

[10] The Microarray Informatics Team at the EBI. ArrayEpress. `www.ebi.ac.uk/arrayexpress`, 2007.

[11] M. Barberis, E. Klipp, M. Vanoni, and L. Alberghina. Cell size at S phase initiation: an emergent property of the G1/S network. *PLoS Comput Biol*, 3(4):e64, 2007.

[12] T. Barrett, T.O. Suzek, D.B. Troup, S.E. Wilhite, W.C. Ngau, P. Ledoux, D. Rudnev, A.E. Lash, W. Fujibuchi, and R. Edgar. NCBI GEO: mining millions of expression profiles database and tools. *Nucleic Acids Res*, 33(suppl 1):D562–566, 2005.

[13] A. Belle, A. Tanay, L. Bitincka, R. Shamir, and E.K. O'Shea. Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences*, 103(35):13004, 2006.

[14] S.J. Benkovic and S. Hammes-Schiffer. A Perspective on Enzyme Catalysis. *Science*, 301(5637):1196–1202, 2003.

[15] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. GenBank. *Nucleic Acids Research*, 35(Database issue):D21, 2007.

[16] A. Beyer, J. Hollunder, H.P. Nasheuer, and T. Wilhelm. Post-transcriptional Expression Regulation in the Yeast Saccharomyces cerevisiae on a Genomic Scale. *Molecular & Cellular Proteomics*, 3(11):1083–1092, 2004.

[17] U.S. Bhalla and R. Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–7, 1999.

[18] International Union Of Biochemistry and Molecular Biology (IUBMB). IUBMB nomenclature home page.
`www.chem.qmul.ac.uk/iubmb/`, 2007.

[19] N. Bluthgen and H. Herzel. How robust are switches in intracellular signaling cascades? *J Theor Biol*, 225(3):293–300, 2003.

[20] B. Boeckmann, M.C. Blatter, L. Famiglietti, U. Hinz, L. Lane, B. Roechert, and A. Bairoch. Protein variety and functional diversity: Swiss-Prot annotation in its biological context Un gène, plusieurs protéines: l'annotation de Swiss-Prot dans le contexte biologique. *Comptes Rendus Biologies*, 328(10-11):882–899, 2005.

[21] S. Borger, W. Liebermeister, and E. Klipp. Distribution of a bifurcation parameter in a genetic network with uncertain parameters. In *Proceedings of the 4th workshop on computation of biochemical pathways and genetic networks*, pages 95–101. Logos-Verlag, Berlin, 2005.

[22] S. Borger, W. Liebermeister, and E. Klipp. Prediction of enzyme kinetic parameters based on statistical learning. *Genome*, 17(1):80–7, 2006.

[23] S. Borger, J. Uhlendorf, A. Helbig, and W. Liebermeister. Integration of Enzyme Kinetic Data from Various Sources. *In Silico Biology*, 7:73–79, 2007.

[24] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, and H.C. et al. Causton. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, 29(4):365–71, 2001.

[25] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G.G. Lara, et al. ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 31(1):68–71, 2003.

[26] A. Brazma, M. Krestyaninova, U. Sarkans, et al. Standards for systems biology. *Nat Rev Genet*, 7(8):593–605, 2006.

[27] P.P.S. Chen. The Entity-Relationship Model-Toward a Unified View of Data. *ACM Transactions on Database Systems*, 1(1):9–36, 1976.

[28] J.M. Cherry, C. Adler, C. Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, et al. SGD: Saccharomyces Genome Database. *Nucleic Acids Research*, 26(1):73–79, 1998.

[29] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, PO Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282(5389):699–705, 1998.

[30] Andrea Ciliberto, Fabrizio Capuani, and John J Tyson. Modeling networks of coupled enzymatic reactions using the total quasi-steady state approximation. *PLoS Comput Biol*, 3(3):e45, 2007.

[31] STRENDA Commission.
`www.strenda.org`, 2007.

[32] Gene Ontology Consortium. Gene Ontology.
`www.geneontology.org/`, 2007.

[33] Gene Ontology Consortium. Saccharomyces Genome Database (SGD) - Progress Report, Gene Ontology Consortium Meeting, Cambridge.
`www.geneontology.org/minutes/20070108_Progress-Reports/SGD-Jan2007.doc`, Jan 2007.

[34] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004.

[35] Open Biomedical Ontologies consortium. Open Biological Ontologies.
`obo.sourceforge.net`, 2007.

[36] The UniProt Consortium. UniProtKB Protein Knowledgebase.
`beta.uniprot.org`, 2007.

[37] The UniProt Consortium. UniProtKB/Swiss-Prot Release 54.3.
`www.expasy.org/sprot/relnotes/relstat.html`, 12 Oct 2007.

[38] The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 35(Database issue):D193–D197, 2007.

[39] J.L. DeRisi, V.R. Iyer, and P.O. Brown. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, 278(5338):680, 1997.

[40] P. Durrens and D.J. Sherman. A systematic nomenclature of chromosomal elements for hemiascomycete yeasts. *Yeast*, 22(5):337–342, 2005.

[41] W. Ebeling and I.M. Sokolov. *Statistical Thermodynamics and Stochastic Theory of Nonequilibrium Systems*. World Scientific, 2005.

[42] European Bioinformatics Institute (EBI). Chemical Entities of Biological Interest (ChEBI).
`www.ebi.ac.uk/chebi/`, 2007.

[43] European Bioinformatics Institute (EBI). MIRIAM Resources.
`www.ebi.ac.uk//miriam/`, 2007.

[44] E.Z. Eisenmesser, D.A. Bosco, M. Akke, and D. Kern. Enzyme Dynamics During Catalysis. *Science*, 295(5559):1520–1523, 2002.

[45] M.B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–8, 2000.

[46] C.P. Fall, E.S. Marland, J.M. Wagner, and J.J. Tyson. *Computational Cell Biology*. Springer, 2005.

[47] D. Fell. *Understanding the control of metabolism*. Portland Press Miami, 1997.

[48] E. Fischer and U. Sauer. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. *Nature Genetics*, 37:636–640, 2005.

[49] National Center for Biotechnology Information (NCBI). Clusters of Orthologous Groups of proteins (COGs).
`www.ncbi.nlm.nih.gov/COG/`, 2007.

[50] National Center for Biotechnology Information (NCBI). GEO (Gene Expression Omnibus).
`www.ncbi.nlm.nih.gov/projects/geo/`, 2007.

[51] National Center for Biotechnology Information (NCBI). NCBI Entrez Genome Project Database.
`www.ncbi.nlm.nih.gov/genomes/leuks.cgi`, 15 Oct 2005.

[52] National Center for Biotechnology Information (NCBI). NCBI Entrez Genome Project Database.
`www.ncbi.nlm.nih.gov/genomes/lproks.cgi`, 15 Oct 2005.

[53] National Center for Biotechnology Information (NCBI). NCBI Entrez Genome Project Database.
`www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj`, Oct 2005.

[54] National Center for Biotechnology Information (NCBI). NCBI Reference Sequence (RefSeq).
`www.ncbi.nlm.nih.gov/RefSeq/`, 2007.

[55] National Center for Biotechnology Information (NCBI). NCBI Entrez Taxonomy.
`www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy`, 2007.

[56] J. Forster, I. Famili, P. Fu, B.O. Palsson, and J. Nielsen. Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network. *Genome Research*, 13(2):244, 2003.

[57] C. Francke, R.J. Siezen, and B. Teusink. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13(11):550–558, 2005.

[58] S. Ghaemmaghami, W.K. Huh, K. Bower, R.W. Howson, A. Belle, N. Dephoure, E.K. O'Shea, and J.S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–741, 2003.

[59] D.T. Gillespie. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *J. Comput. Phys*, 22(4):403–434, 1976.

[60] D.T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188(1-3):404–425, 1992.

[61] D.T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001.

[62] R.N. Goldberg. Thermodynamics of enzyme-catalyzed reactions–a database for quantitative biochemistry. *Bioinformatics*, 20(16):2874–2877, 2004.

[63] R.N. Goldberg, Y.B. Tewari, and T.N. Bhat. Thermodynamics of Enzyme-Catalyzed Reactions database (TECRDB).
`xpdb.nist.gov/enzyme_thermodynamics/`, 2007.

[64] A. Goldbeter. Computational approaches to cellular rhythms. *Nature*, 420(6912): 238–245, 2002.

[65] A. Goldbeter and D.E. Koshland. An amplified sensitivity arising from covalent modification in biological systems. *Proc Natl Acad Sci USA*, 78(11):6840–4, 1981.

[66] A. Goldbeter and D.E. Koshland. Ultrasensitivity in biochemical systems controlled by covalent modification. Interplay between zero-order and multistep effects. *Journal of Biological Chemistry*, 259(23):14441–14447, 1984.

[67] D. Greenbaum, R. Jansen, and M. Gerstein. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics*, 18 (4):585–596, 2002.

[68] D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol*, 4(9): 117, 2003.

[69] N. Gunnarsson, A. Eliasson, and J. Nielsen. Control of fluxes towards antibiotics and the role of primary metabolism in production of antibiotics. *Advances in biochemical engineering, biotechnology*, 88:137–178, 2004.

[70] J. Hakenberg, S. Schmeier, A. Kowald, E. Klipp, and U. Leser. Finding kinetic parameters using text mining. *Omics: A Journal of Integrative Biology*, 8(2): 131–152, 2004.

[71] K. Hartmann and D. Schomburg. GibbsPredictor: Predicting Gibbs energies from molecular structures. *Bioinformatics*, 2006. submitted.

[72] D.T. Haynie. *Biological Thermodynamics*. Cambridge University Press, 2001.

[73] F.C.P. Holstege, E.G. Jennings, J.J. Wyrick, T.I. Lee, C.J. Hengartner, M.R. Green, T.R. Golub, E.S. Lander, and R.A. Young. Dissecting the Regulatory Circuitry of a Eukaryotic Genome. *Cell*, 95(5):717–728, 1998.

[74] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, A.P. Arkin, B.J. Bornstein, D. Bray, A. Cornish-Bowden, A.A. Cuellar, S. Dronov, E.D. Gilles, M. Ginkel, V. Gor, I.I. Goryanin, W.J. Hedley, T.C. Hodgman, J.H. Hofmeyr, P.J. Hunter, N.S. Juty, J.L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L.M. Loew, D. Lucio, P. Mendes, E. Minch, E.D. Mjolsness, Y. Nakayama, M.R. Nelson, P.F. Nielsen, T. Sakurada, J.C. Schaff, B.E. Shapiro, T.S. Shimizu, H.D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language(SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[75] W.K. Huh, J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman, and E.K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–691, 2003.

[76] Munich information center for protein sequences. Comprehensive Yeast Genome Database (CYDG).
`mips.gsf.de/genre/proj/yeast`, 2007.

[77] International BioModels.net Initiative.
`biomodels.net`, 2007.

[78] International BioModels.net Initiative. BioModels Database.
`www.ebi.ac.uk/biomodels/`, 2007.

[79] International Nucleotide Sequence Database Collaboration (INSDC).
`www.insdc.org/`, 2007.

[80] The International Nucleotide Sequence Database Collaboration (INSDC). EMBL Nucleotide Sequence Database.
`www.ebi.ac.uk/embl/`, 2007.

[81] SRI International. BioCyc database collection.
`www.biocyc.org/`, 2007.

[82] O.N. Jensen. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol*, 8(1): 33–41, 2004.

[83] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[84] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30(1):42–46, 2002.

[85] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(90001):277–280, 2004.

[86] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34:D354–D357, 2006.

[87] P.D. Karp, C.A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahrén, S. Tsoka, N. Darzentas, V. Kunin, N. López-Bigas, et al. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 33(19):6083–6089, 2005.

[88] B.N. Kholodenko. Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades, 2000.

[89] B.N. Kholodenko. Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell Biol*, 7(3):165–176, 2006.

[90] P.M. Kim and B. Tidor. Limitations of Quantitative Gene Regulation Models: A Case Study. *Genome Research*, 13(11):2391–2395, 2003.

[91] E. Klipp, W. Liebermeister, and C. Wierling. Inferring dynamic properties of biochemical reaction networks from structural knowledge. *Genome Inform Ser Workshop Genome Inform*, 15(1):125–37, 2004.

[92] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice: concepts, implementation and application*. Wiley-VCH, 2005.

[93] E. Klipp, B. Nordlander, R. Krueger, P. Gennemark, and S. Hohmann. Integrative model of the response of yeast to osmotic shock. *Nature Biotechnology*, 23(8):975–982, 2005.

[94] E. Klipp, W. Liebermeister, A. Helbig, A. Kowald, and J. Schaber. Systems biology standards – the community speaks. *Nature Biotechnology*, 25:390–391, 2007.

[95] C. J. Krieger, P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee, and P. D. Karp. Metacyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, 32 Database issue:D438–42, 2004.

[96] L. Kuepfer, U. Sauer, and L.M. Blank. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome Research*, 15(10):1421, 2005.

[97] T. Kulikova, R. Akhtar, P. Aldebert, N. Althorpe, M. Andersson, A. Baldwin, K. Bates, S. Bhattacharyya, L. Bower, P. Browne, et al. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Research*, 35(Database issue):D16, 2007.

[98] T.G. Kurtz. The Relationship between Stochastic and Deterministic Models for Chemical Reactions. *The Journal of Chemical Physics*, 57:2976–2978, 1972.

[99] Kanehisa Laboratories. KEGG: Kyoto Encyclopedia of Genes and Genomes. `www.genome.ad.jp/kegg/`, 2007.

[100] N. Le Novère, A. Finney, M. Hucka, U.S. Bhalla, F. Campagne, J. Collado-Vides, E.J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B Shapiro, J.L. Snoep, H.D. Spence, and B.L. Wanner. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*, 23:1509–1515, 2005.

[101] V. Leskovac. *Comprehensive Enzyme Kinetics*. Springer, 2003.

[102] W. Liebermeister and E. Klipp. Biochemical networks with uncertain parameters. *Systems Biology, IEE*, 152(3):97–107, 2005.

[103] W. Liebermeister and E. Klipp. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theoretical Biology and Medical Modelling*, 3(1):41, 2006.

[104] W. Liebermeister and E. Klipp. Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. *Theoretical Biology and Medical Modelling*, 3(1):42, 2006.

[105] Eli Lilly, Company, and NIH Chemical Genomics Center. Assay Guidance Manual Version 4.1.
`www.ncgc.nih.gov/guidance/manual_toc.html`, 2005.

[106] X. Mao, T. Cai, J.G. Olyarchuk, and L. Wei. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, 21(19):3787, 2005.

[107] R. Milo, P. Jorgensen, M. Springer, and G. Weber. BioNumbers, The Database to Useful Biological Numbers.
`bionumbers.hms.harvard.edu/`, 2007.

[108] C.J. Mungall. Obol: integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics*, 5(6-7):509–520, 2004.

[109] A. Nanchen, A. Schicker, and U. Sauer. Nonlinear Dependency of Intracellular Fluxes on Growth Rate in Miniaturized Continuous Cultures of Escherichia coli. *Applied and Environmental Microbiology*, 72(2):1164–1172, 2006.

[110] D. Noble. Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations. *Nature*, 188(495-497):111, 1960.

[111] D. Noble. *The Music of Life: Biology Beyond the Genome*. Oxford University Press, 2006.

[112] National Institutes of Health (NIH). GenBank.
`www.ncbi.nlm.gov/Genbank/`, 2007.

[113] U.S National Library of Medecine (NLM) and National Institutes of Health (NIH). PubMed.
`www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed`, 2007.

[114] Internation Union of Pure and Applied Chemistry (IUPAC). IUPAC Nomenclature.
`www.chem.qmul.ac.uk/iupac/`, 2007.

[115] Internation Union of Pure and Applied Chemistry (IUPAC). IUPAC Recommendations for nomenclature and tables in biochemical thermodynamics.
`www.chem.qmul.ac.uk/iupac/thermod/`, 2007.

[116] Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature.
`www.chem.qmul.ac.uk/iubmb/enzyme/`, 2007.

[117] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.

[118] S.G. Oliver et al. From DNA sequence to biological function. *Nature*, 379(6566): 597–600, 1996.

[119] I. Oppenheim, KE Shuler, and GH Weiss. Stochastic and Deterministic Formulation of Chemical Rate Equations. *The Journal of Chemical Physics*, 50:460–466, 1969.

[120] O'Shea and Weissman labs at UCSF. Yeast GFP fusion database.
`yeastgfp.yeastgenome.org`, 2007.

[121] R. Overbeek, T. Begley, R.M. Butler, J.V. Choudhuri, H.Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, et al. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucleic Acids Research*, 33(17):5691, 2005.

[122] L. Pena-Castillo and T.R. Hughes. Why Are There Still Over 1000 Uncharacterized Yeast Genes? *Genetics*, 176(1):7, 2007.

[123] SGD project. Saccharomyces Genome Database. `www.yeastgenome.org/`, 2007.

[124] The Reactome Project. Reactome - a curated knowledgebase of biological pathways. `www.reactome.org/`, 2007.

[125] K.D. Pruitt, T. Tatusova, and D.R. Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Database Issue):D501, 2005.

[126] D.L. Purich and R.D. Allison. *Handbook of biochemical kinetics*. Academic Press New York, 2000.

[127] C.V. Rao and A.P. Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *The Journal of Chemical Physics*, 118:4999, 2003.

[128] M. Rizzi, M. Baltes, U. Theobald, and M. Reuss. In vivo analysis of metabolic dynamics in Saccharomyces cerevisiae: II. Mathematical model. *Biotechnology and Bioengineering*, 55(4):592–608, 1997.

[129] U. Sauer. Metabolic networks in motion: 13 C-based flux analysis. *Mol Syst Biol*, 2:62, 2006.

[130] I. Schomburg, A. Chang, and D. Schomburg. BRENDA, enzyme data and metabolic information. *Nucleic acids research*, 30(1):47–49, 2002.

[131] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic acids research*, 32:431–433, 2004.

[132] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4: 41, 2003.

[133] C.F. Taylor, N.W. Paton, K.S. Lilley, P.A. Binz, R.K. Julian Jr, A.R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E.W. Deutsch, M.J. Dunn, A.J.R. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T.A. Neubert, S.D. Patterson,

P. Ping, S.L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T.M. Vondriska, J.P. Whitelegge, M.R. Wilkins, I. Xenarios, J.R. Yates 3rd, and H. Hermjakob. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol*, 25(8):887–93, 2007.

[134] BRENDA team. BRENDA, The Comprehensive Enzyme Information S. `www.brenda-enzymes.info`, 2007.

[135] The SMBL Team. Systems Biology Markup Language (SBML). `www.sbml.org`, 2007.

[136] B. Teusink, J. Passarge, C.A. Reijenga, E. Esgalhado, C.C. van der Weijden, M. Schepper, M.C. Walsh, B.M. Bakker, K. van Dam, H.V. Westerhoff, et al. Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem*, 267(17):5313–5329, 2000.

[137] M. Thorsen, G. Lagniel, E. Kristiansson, C. Junot, O. Nerman, J. Labarre, and M.J. Tamas. Quantitative transcriptome, proteome, and sulfur metabolite profiling of the Saccharomyces cerevisiae response to arsenite. *Physiological Genomics*, 30 (1):35, 2007.

[138] J.J. Tyson, K.C. Chen, and B. Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol*, 15(2):221–231, 2003.

[139] A. Varshavsky. The N-End Rule: Functions, Mysteries, Uses. *Proc Natl Acad Sci USA*, 93(22):12142–12149, 1996.

[140] I. Vastrik, P. D'Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, et al. Reactome: a knowledge base of biologic pathways and processes. *Genome Biology*, 8(3):R39, 2007.

[141] J.M.G. Vilar, C.C. Guet, and S. Leibler. Modeling Network Dynamics: The lac Operon, a Case Study. *The Journal of Cell Biology*, 161(3):471–476, 2003.

[142] E.C. Webb. *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press, 1992.

[143] EC Webb. Enzyme nomenclature: a personal retrospective. *FASEB J*, 7(12): 1192–4, 1993.

[144] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35 (Database issue):D5, 2007.

[145] D.S. Wishart, D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, et al. HMDB: the Human Metabolome Database. *Nucleic Acids Research*, 35(Database issue):D521, 2007.

[146] L. Wodicka, H. Dong, M. Mittmann, M.H. Ho, and D.J. Lockhart. Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat Biotech*, 15(13):1359–1367, 1997.

[147] W. Xiong. A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature*, 426:460–465, 2003.

List of Figures

List of Tables

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 28.01.2008 Simon Borger

Acknowledgements

I am especially grateful to my supervisor Prof Edda Klipp and my colleague Dr Wolfram Liebermeister.

I also thank Dr Axel Kowald, Jannis Uhlendorf and Anselm Helbig for help in the acquisition of data for the database. Furthermore, Anselm Helbig was of great help in improving the database. Working with him taught me a lot in programming.

The Graduate Program Berlin "Dynamics and Evolution of Cellular and Macromolecular Processes" (the predecessor of the International Research Training Group “Genomics and Systems Biology of Molecular Networks”) supported my PhD thesis from Januar 2004 to December 2006. In 2007 until October I was supported by the ENFIN network of excellence (Enabling Systems Biology) LSHG-CT-2005-518254.